A NOVEL BIOINFORMATIC APPROACH FOR COMPREHENSIVE GENOME SCALE ANALYSIS IDENTIFIES KEY REGULATORS OF MACROPHAGE ACTIVATION

Samuel Katz

Supervisors:

Prof. Clare E. Bryant

Dr. Iain D.C. Fraser

University of Cambridge

Fitzwilliam College

This dissertation is submitted for the degree of Doctor of Philosophy.

September 2019

Declaration and Statement of Length

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

The dissertation does not exceed the prescribed word limit for the relevant Degree Committee.

Samuel Katz

September 2019

A Novel Bioinformatic Approach for Comprehensive Genome Scale Analysis Identifies Key Regulators of Macrophage Activation

Samuel Katz

Summary

The initiation of inflammatory cytokine transcription by bacterial ligands is a central mechanism by which the immune system activates its first line of defense. Macrophage activation by the Toll-like Receptor 4 (TLR4) pathway is initiated with receptor binding of lipopolysaccharides (LPS) and culminates in a large-scale transcriptional response of the inflammatory gene program. Advancements in genome-wide screening technologies have made it possible to interrogate the regulatory landscape of signaling pathways such as those activated by TLR4. Utilizing these high-throughput methods for the comprehensive characterization of pathway components, particularly for regulators that are involved in critical cellular processes such as transcription and translation, however, requires an approach that goes beyond the top scoring and previously characterized hits of genome-scale studies. To address this challenge, I developed the Throughput Ranking by Iterative Analysis of Genomic Enrichment (TRIAGE) method, a bioinformatic analysis model that facilitates the comprehensive identification of likely regulators by iterative sampling of pathway and network databases. I validated the TRIAGE approach by analyzing three previously published genome-wide studies of regulators of early HIV infection and viral transcription. Analysis by TRIAGE showed significantly increased overlap and identified shared novel targets across the three studies. I further developed the TRIAGE analysis method as a globally accessible web-based resource. Applying TRIAGE analysis to three genome-scale studies of LPS treatment in macrophages of mouse and human cell lines, I identified an enrichment for regulators relating to alternative splicing and protein degradation. Using short read and long read RNA-seq of ligand-stimulated macrophages I further characterized the broad transcriptional variation induced by the LPS response and the novel and known transcript variants that define different macrophage activation states. These findings define an approach for comprehensive unbiased discovery of signaling pathway regulators from genome-scale datasets and suggest a model of macrophage activation involving proteasomal removal of negative regulators and remodeling of the macrophage state via a transcriptional shift in splice variant dynamics.

Publications arising from this thesis

Sun, J., Katz, S., Dutta, B., Wang, Z., & Fraser, I. D. (2017). Genome-wide siRNA screen of genes regulating the LPS-induced TNF-alpha response in human macrophages. *Scientific Data, 4*, 170007. doi:10.1038/sdata.2017.7

Li, N., Katz, S., Dutta, B., Benet, Z. L., Sun, J., & Fraser, I. D. (2017). Genome-wide siRNA screen of genes regulating the LPS-induced NF-κB and TNF-α responses in mouse macrophages. *Scientific Data*, *4*, 170008. doi:10.1038/sdata.2017.8

Acknowledgements

The world so rarely lets us in, let's praise the luck vista when it does.
-Linda Gregerson, Prodigal: New and Selected Poems, 1976 to 2014

At the start of my work that would become this document I came across the above words in a poem describing the findings by Brenner, Horvitz, and Sulston of how the processes of controlled cell death are essential for the development of life. From the day I read them to this moment now as I write this sentence these words have conveyed the luck, joy, and awe I feel about my utter privilege of being able to ask questions of the biological world and the possibility of one day being let in on so elegant an answer. I am filled with gratitude towards the people who have shared and made possible this path that I am on.

My family who have shared all the different sides of this journey with me. My sister Pessy who understands better than anyone what the highs and lows of this process mean for me. My sister Chavali with whom I share our own unique joys of what making a new place your home means. My brother-in-law Michael whom I have known for the same amount of time as the work in this thesis, and honestly, both feel like they were always there. My brothers Avi and Gavi who ask me the best questions about my work. Sharry, who prepared my family for what an undertaking this document can be. My father, Ysoscher Katz, Your Heuristic Worldview Happened. And who for years told me that he will only understand the first four words of my thesis ("this thesis is about") but understands everything that is packed in the space between the words for me. My mother, Esther Katz, who in a world so far away still sees me doing her God's work. My grandmother, Gita Katz who never misses a chance for a good and real conversation. Two people in my family only got to be on part of this journey with me, my grandfather Leiby Katz who left this world soon after I started this work and my niece Adi Grummer who came into it just as I was completing it. Both of whom are my reminders for how truly awesome the gifts of life, health, and curiosity are.

Many friends lifted me up through this journey. Emily Lees, Alison Clasper, and Geoff Eichhorn; thank you for the best of my Cambridge days. My dear friends in New York who always have a pair of eyes to lend me when I forget how to see myself, Malky Brizel Lipshitz, Brooke Sable, and Shmuly Horowitz; thank you for knowing where I come from and seeing where I am going. Marlon Kazmierczak for reminding me of the life that exists outside of this world. Gianmarcco Raddi, who will show up late to read this, but he will show up. Solomon Feuerwerker for sharing with me the joy of seeing what our early dreams became. Anita Gola,

guided me through the hilarity and stress that this process can be and helped meet the finish line with sanity. Nafisa Waziri for the *santi* of it all. Brian Graves for understanding how it can be like that.

Danette Hargraves has been a mentor to me since the days when I was working on my high school equivalency degree and has continued to cheer me on through years of undergraduate education, jobs, applications, and this work. Thank you for your wisdom and guidance at every step, and for the words that got me through at the lowest point in this journey.

Dr. Beth Moore gave me my first lab project and showed me the joy of diligent research. It was in her lab that I was first exposed to the fascinating questions of immunology and the wonderful complexity of the macrophage.

The NIH-Oxford-Cambridge Scholars program made this work possible. My thanks to the admin team, the advisory board, and the International Biomedical Research Alliance.

My gratitude to the members of the Bryant Lab in Cambridge. Lee (Hoppy) Hopkins, who were it not for his Welsh pride I would have mistaken for a long-lost brother from New York. Pani Tourlomousis, Charlotte Macleod, Milton Pereira, and especially Zsofi Digby, we did get it in the end, like I kept saying.

The Fraser group at the NIH made every detail of the work in this thesis possible. Jing Sun, Sharat Jacob Vayttaden, Sinu John, Orna Ernst Rabinovich, Clinton Bradfield, Jonathan Liang, and Mike Dorrington. And former members Nicolas Lounsbury, Bin Lin, and Kyle Webb. All of my work in here stands on the shoulders of what they have done and taught me.

I am incredibly grateful to Dr. Jian Song who expanded what the TRIAGE project could be and helped me get it to where it is now.

Finally, and most importantly, my deepest gratitude goes to my mentors of this work.

Clare Bryant gave me the space in her lab to explore in a new way a biological system she knows better than most. Clare modeled and taught me a healthy dose of scientific skepticism that formed so much of how my thoughts were shaped in this process.

Iain Fraser gave me time, resources, and support to discover the questions I was passionate about and guided me as I grew in my abilities to pursue it. Living through the logistics of what this project took, both inside and outside of the lab, would not have been possible without his support. Iain never accepted my arguments about not me being innately good at certain kinds of research techniques by reminding me that scientific intuition is learned. Iain taught me that doing something doesn't mean doing it all by yourself. I am overwhelmed with appreciation that I got to develop my early scientific thinking under his teaching, expertise, and scientific sensibility. No praise could adequately communicate my gratitude.

Dedication

On April 5th, 2008 I made a choice that became my first step towards reaching this milestone.

It is for that day and what it took to make that choice that I dedicate this thesis.

Ushy, it has all been worth it.

Samuel Katz

September 30, 2019

Table of Contents

List of Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

List of Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 5

Chapter 7

List of Abbreviations

Abbreviation	Full Form
A3SS	Alternative 3' Splice Site
A5SS	Alternative 5' Splice Site
AIM	Absent In Melanoma
ALR	AIM2-like receptor
API	Application Program Interface
ARDS	Acute Respiratory Distress Syndrome
BMDM	Bone Marrow Derived Macrophages
bp	Base Pair
CARD	Comprehensive Analysis of RNAi Data
CD	Cluster of Differentiation
cGAS	cyclic GMP-AMP synthetase
CLR	C-type lectin receptor
CRISPR	Clustered Regularly Interspersed Palindromic Repeats
CYLD	Cylindromatosis
DAMP	Danger Associated Molecular Pattern
DC	Dendritic Cell
DMSO	dimethyl sulfoxide
DNA	deoxyribonucleic acid
ESE	Exonic Splicing Enhancers
ESS	Exonic splicing silencer
FDR	False Discovery Rate
FET	Fisher Exact Test
FSM	Full Splice Match
GFP	Green Fluorescent Protein
GM-CSF	Granulocyte-Macrophage Colony Stimulating Factor
HDF	Host Deficiency Factors
HEB	Hierarchical Edge Bundling
HGNC	HUGO Gene Nomenclature Committee
HIV	Human Immunodeficiency Virus
hnRNP	Heterogeneous Ribonucleoprotein Particle
IAP	Inhibitor of Apoptosis Protein
IFN	Interferon
IKK	IκB kinase
IL	Interleukin
IPA	Ingenuity Pathway Analysis
IRAK	Interleukin-1 Receptor-Associated Kinase
IRF	interferon regulatory factor
ISE	Intronic Splicing Enhancers
ISM	Incomplete Splice Match
JNK	c-Jun N-terminal kinases

K48	lysine-48
K63	lysine-63
KD	Knockdown
Kdo	deoxy-d-*manno*-oct-ulosonic acid
KEGG	Kyoto Encyclopedia for Genes and Genomes
KO	Knockout
LDH	Lactate Dehydrogenase
LPS	lipopolysaccharides
M-CSF	Macrophage Colony-Stimulating factor
Madrasin	2-((7methoxy-4-methylquinazolin-2-yl)amino)-5,6-dimethylpyrimidin-4(3H)-one RNAsplicing inhibitor
MAPK	Mitogen-Activated Protein Kinase
MBNL1	Muscleblind Like Splicing Regulator 1
MD2	myeloid differentiation factor 2
MDA5	Melanoma-Differentiation-Associated Gene 5
MDS	Myelodysplastic Syndromes
MFI	Mean Fluorescent Intensity
Mo	Monocyte
MXE	Mutually Exclusive Exons
MyD88	Myeloid differentiation primary response 88
NA	Network Analysis
NFκB	Nuclear Factor Kappa-light-chain-enhancer of activated B cells
NIC	Novel In Catalog
NLRs	NOD-like receptors
NNC	Novel Not in Catalog
OST	Oligosacchyraltransferase
PA	Pathway Analysis
PAMP	Pathogen Associated Molecular Pattern
PBMC	Peripheral Blood Mononuclear Cell
PMA	Phorbol Myristate Acetate
PPI	Predicted Protein Interactions
PRR	Pattern Recognition Receptor
qPCR	quantitative Polymerase Chain Reaction
RI	Retained Intron
RIG-I	Retinoic acid-Inducible Gene I
RLR	RIG-I-like receptor
rMATS	replicate Multivariate Analysis of Transcript Splicing
RNA	ribonucleic acid
RNA-seq	RNA-sequence
RNAi	RNA interference
SE	Skipped Exon
SF3B	Splicing Factor 3b Subunit
SILAC	Stable Isotope Labeling with Amino Acids in Cell Culture
siRNA	short interfering RNA

SQANTI	Structural and Quality Annotation of Novel Transcript Isoforms
SR	Serine-Arginine
SRSF	Serine and Arginine Rich Splicing Factor
STING	Stimulator of Interferon Genes
STRING	Search Tool for the Retrieval of Interacting Genes
TAB	TAK1-binding protein
TAG	TRAM Adaptor Molecule with GOLD Domain
TAK	TGFβ-activated kinase
TICAM	TIR-containing adapter molecule
TIR	Toll/IL-1 receptor
TIRAP	TIR Domain Containing Adaptor Protein
TLR	Toll-like Receptor
TNFα	Tumor Necrosis Factor alpha
TOLLIP	Toll-Interacting Protein
TRAF	TNF-receptor associated factor
TRAM	TRIF-Related Adaptor Molecule
TRIAGE	Throughput Ranking by Iterative Analysis of Genomic Enrichment
TRIF	TIR Domain-containing Adaptor-inducing Interferon-β
U2AF	U2 Small Nuclear RNA Auxiliary Factor
UBL5	Ubiquitin-like protein 5
WT	Wildtype

A NOVEL BIOINFORMATIC APPROACH FOR COMPREHENSIVE GENOME SCALE ANALYSIS IDENTIFIES KEY REGULATORS OF MACROPHAGE ACTIVATION

1 Introduction

1.1 Innate immunity, macrophages, and inflammation

1.1.1 The innate immune response

An immune response is when a subset of cells within an organism synchronize their activity to eliminate or control a perceived threat. Immune responses vary in time, specificity, and memory, with different immune responses falling under different combinations of these criteria. A prevailing model of the mammalian immune system categorizes immune response types into two groups, adaptive and innate (C. A. Janeway et al., 1996). Adaptive is defined as immune responses that develop over time with increased specificity towards the targets of danger. Adaptive immune responses also maintain recognition and responsiveness after the challenge has been eliminated. By contrast, innate immune activity is generalized as a response that cells can mount quickly and effectively but whose impact is less adapted to the specific challenge the organism encounters. In the innate immunity model, once the challenge has been controlled, the cells and tissue return to homeostasis with a secondary encounter of the challenge restarting the response same as it did with the first encounter.

Many of the tenants of the adaptive vs. innate model of immunity have been challenged since its introduction. Cells such as dendritic cells have been shown to engage and cross between both branches of the response (Hammad & Lambrecht, 2011). Certain innate immune responses have been shown to have an element of training that changes the response with repeated exposure (Netea et al., 2016). The rapid responsiveness aspect of the innate immunity model, however, has remained central to our understanding of how mammalian immune systems operate. Organisms have had to evolve dedicated systems that can in a short time mount effective immune responses (Kimbrell & Beutler, 2001). This evolution had to happen in parallel with a system that prevents the engagement of these responses too frivolously.

Failure of that system leads to excessive or sterile inflammation which are the central pathogenesis of autoinflammatory and autoimmune diseases (Arakelyan et al., 2017).

1.1.2 Macrophages: diversity & polarity

One of the key components of innate immunity are macrophage cells. Macrophages survey their environment, respond to challenges, and repair tissue following an immune response. From the beginning of their discovery, two different behaviors were observed with macrophages. Some observed macrophages were fixed to certain tissue environments while others were seen to traffic through the blood stream (Gordon, 2016). This has led to a current understanding that categorizes macrophages into two groups, those that are tissue resident and those which circulate through the blood and are recruited to sites of challenge by the sensing of microorganisms or homing chemokines and cytokines.

Early discoveries found that macrophages were derived from bone marrow cells (van Furth & Cohn, 1968). A proposed model for what differentiated the fixed versus migratory macrophages suggested that both macrophage types originated from bone marrow monocytes with tissue specific macrophages having low levels of the Ly6C marker and circulating macrophages having $Ly6C^{hi}$ (Geissmann et al., 2003; Lawrence & Natoli, 2011). Later work by a number of groups found that macrophages that are specific to certain tissue derive from fetal monocytes that originate in the yolk sac or fetal liver (Guilliams et al., 2013; Schulz et al., 2012; Yona et al., 2013). In this model, tissue resident macrophages develop from fetal precursors and remain to survey the same tissue for the lifetime of the host, while recruited macrophages develop from bone marrow derived monocytes and circulate through the blood patrolling for sites or causes of immune challenge (Yona et al., 2013). Though the embryonic origin for some populations of self-sustaining tissue resident macrophages is now widely

accepted, emerging work suggests that tissue resident macrophages consist of diverse subsets of both embryonic and bone marrow origin (Epelman et al., 2014; Guilliams et al., 2018).

Functionally, macrophages possess a range of seemingly opposing capacities. In the encounter of an immune challenge, macrophages activate microbicidal and apoptotic programs, whereas in the context of injury and in the aftermath of an immune response, they can promote tissue repair and a return to homeostasis. The two activation states can be induced *in vitro* by different stimuli. Treatment of macrophages by lipopolysaccharides (LPS) or interferon (IFN)-γ induces pro-inflammatory activity (Dalton et al., 1993), while treatment by interleukin (IL) 4 or IL13 leads to the expression of anti-inflammatory cytokines (Stein et al., 1992). Early models defined this diversity of activation states into separate groups of classical activation (pro-inflammatory and pro-death) for the LPS/IFN-γ treated macrophages and alternative activation (anti-inflammatory and pro-healing) for the IL4/IL13 treated macrophages (Edwards et al., 2006). This model was then replaced by a paradigm to parallel T cell activation and the model of Th1/Th2 responses. This saw the two activation paths in macrophages as part of a polarization of the cell to match the T cell states and defined the polarities as M1 (pro-inflammatory) and M2 (anti-inflammatory) (Mills et al., 2000).

Subsequent work, however, has challenged the idea that macrophage activation is determined by divergent polarization and limited to the binary states of pro- and anti-inflammatory. Work done in human cells found that an overly simplified M1/M2 polarity did not represent the full spectrum of observed functions, and that activation in different contexts can lead to more varied functions across the range of these macrophages polarities (Mosser & Edwards, 2008; Murray & Wynn, 2011; Xue et al., 2014).

1.1.3 The inflammatory response

Though it is context and stimulus dependent, the most studied function of macrophages is their release of inflammatory cytokines as a pathogen fighting and immune protective response (Arango Duque & Descoteaux, 2014). The range of inflammatory cytokines secreted by macrophages include, but are not limited to, TNFα (B. Beutler et al., 1985), IL1 (Carmi et al., 2009), IL6 (Hurst et al., 2001), IL12 (Trinchieri, 1994), and type I IFNs (Decker et al., 2005). These cytokines, together with a collection of chemokines, further recruit additional immune cells to sites of infection or tissue damage, induce apoptosis of infected cells, trigger the expression of anti-microbial proteins in targeted cells, and induce phagocytosis of pathogenic debris. Following the clearance of the immunological challenge, macrophages secrete a set of anti-inflammatory cytokines that regulate and resolve their pro-inflammatory activity (Chadban et al., 1998; Xiao et al., 2002).

1.2 The PRR-ligand model; Toll-like receptor 4 and lipopolysaccharide

1.2.1 PRRs

As stationed or circulating surveyors of immunity, macrophages require a mechanism by which to selectively sense the presence of immune challenges. Since they lack the capacity of adaptive immune cells for clonal expansion and variability, it was not initially appreciated that macrophages can discriminate between different immunological threats and mount specific responses. Janeway was first to propose a model of innate immune cell sensing that relied on a set of a specific receptors termed Pattern Recognition Receptors (PRRs) (C. A. Janeway, Jr., 1989). Subsequent work has identified numerous types of PRRs that are localized to different segments of the cell and whose receptors are specific for different groups of immune challenges. The first such a PRR was the family of Toll-like receptors (TLRs) (Hoffmann, 2003; Medzhitov et al., 1997) (described in more detail in section 1.2.3). TLRs are

transmembrane proteins that detect specific molecular patterns of microbes (see section 1.2.2 on PAMPs) and initiate transcription of pro-inflammatory cytokines and type-I IFNs. Additional PRRs were subsequently identified with variations in their localization, pattern recognition, and downstream response induction. C-type lectin receptors (CLRs) are transmembrane receptors like TLRs but are relegated to the plasma membrane and predominately recognize fungal infections (Netea et al., 2006). NOD-like receptors (NLRs) recognize microbe motifs, but in contrast to TLRs are localized in the cytoplasm (Fritz et al., 2006). A series of cytoplasmic localized receptors which recognize nucleic acids were also identified. RIG-I-like receptors (RLRs) (Hornung et al., 2006; Yoneyama et al., 2004) (including the melanoma-differentiation-associated gene 5 (MDA5) (Kato et al., 2006)), absent in melanoma 2 (AIM2)-like receptors (ALRs) (Bürckstümmer et al., 2009; Fernandes-Alnemri et al., 2009; Hornung et al., 2009), and cyclic GMP-AMP synthetase (cGAS)-stimulator of interferon genes (STING) (Ishikawa & Barber, 2008; Ishikawa et al., 2009; X. Li et al., 2013; L. Sun et al., 2013). Different cytosolic receptors recognize different forms of nucleotides. RLRs respond predominantly to viral RNA (though it can also respond to exposed bacterial RNA) (Hornung et al., 2006). cGAS-STING responds to cytosolic DNA (Ishikawa et al., 2009). Initiation of inflammatory cytokine transcription and type-I IFNs is also not the only response initiated by PRRs. Some PRRs recruit super molecular complexes to specific locations in the cell. Activation of NLRs and ALRs also activate the inflammasome complex and lead to the maturation of Interleukin (IL)-1 and cell death (Franchi et al., 2012). An additional protein, Caspase-11 was found to be able to recognize intracellular endotoxins and activate a non-canonical inflammasome pathway (Hagar et al., 2013; Kayagaki et al., 2013). The diversity and specificity of these receptors support the original hypothesis by Janeway that the effectors of innate immunity have evolved sets of PRRs that recognize and discriminate between targets. Recent work has added the localization of receptors as additionally critical in

ensuring that the receptor's responsiveness is not activated by the host, in addition to the receptor specificity as proposed by Janeway (Chow et al., 2015).

1.2.2 PAMPs and DAMPs

A necessary corollary to the PRR model proposed by Janeway was that pathogens contain pathogen-associated molecular patterns (PAMPs) and that these patterns are unique to pathogens and absent in the host organism. An additional criterion for PAMPs is that they involve components essential for pathogen function and survival to circumvent evolutionary pressure on the pathogen to easily change these factors (C. A. Janeway, Jr., 1989). This model aligned with Janeway's earlier dictum of innate immunity, that it evolved to discriminate between self and non-self molecules of the host organism (C. A. Janeway, 1992). A contradictory model proposed by Matzinger and later confirmed by others found that even in the absence of a pathogen, signals associated with danger to the host cell or tissue, such as necrosis, can also activate the innate immune system (Matzinger, 1994; Scaffidi et al., 2002; Scheibner et al., 2006). These findings amended the category of triggers for innate immunity to include PAMPs and danger-associated molecular patterns (DAMPs). Further work also suggested the addition of another subcategory of molecular patterns associated with tissue trauma, termed alarmins (Bianchi, 2007). (It has been argued, however, that this category can be considered to fall under the category of DAMPs). Though these categories differ in the central dogma of innate immunity they profess, common to all of them is that the binding of a distinct molecular pattern with a complementary evolved receptor triggers and initiates innate immunity in different contexts and that the diversity of immunological challenges can be discriminated on the basis of specific receptor-ligand interactions.

1.2.3 LPS and TLR4

Predating our understanding of receptors and ligands in innate immunity by almost a century, independent work by Pfeiffer and Centanni found that endotoxins -integral components of the outer membranes of gram-negative bacteria- were sufficient to induce the deleterious effects of bacteria (*Vibrio cholerae* and *Salmonella typhi,* respectively) (Centanni & Bruschettini, 1894; Pfeiffer, 1892). These immunologically active endotoxins were later purified by Boivin, Lüderitz, and Westphal and identified as lipopolysaccharides, or LPS (Boivin & Mesrobeanu, 1935; Lüderitz et al., 1971). The structure of LPS is comprised of three domains, a lipid component that is anchored to the bacterial cell wall called lipid A, a core oligosaccharide bound to lipid A by a 3-deoxy-d-*manno*-oct-ulosonic acid (Kdo) called the core component, followed by a variable chain of polysaccharide called the O-antigen (Figure 1.1A) (Erridge et al., 2002). The O-antigen, which is the outermost structure of LPS, is variable across different species of bacteria and it is how the presence or absence of a specific pathogen in the blood can be determined (i.e. "serological specificity"). The lipid A domain conserves some architecture across species including two phosphate groups and two acyloxyacyl moieties (though position and length can vary from species to species) and it is what gives LPS its endotoxic activity (Alexander & Rietschel, 2001; Raetz & Whitfield, 2002). LPS was also found to induce the expression of Tumor Necrosis Factor (TNF), an inflammatory cytokine, in macrophages (B. A. Beutler et al., 1985). Thus, LPS became the first characterized PAMP.

Identifying the first PRR took a more circuitous path. A *Toll* gene was first identified in *Drosophila* as critical for guiding spatial embryonic development (Hashimoto et al., 1988). Later work found that Toll -along with its ligand, spätzle, and its inhibitor, cactus- also functions as a receptor in the adult *Drosophila* anti-fungal response (Lemaitre et al., 1996). It was also observed that the activation dynamics of Toll in *Drosophila* bore a striking resemblance to the mammalian immune pathway induced through the IL-1 receptor (Gay &

Keith, 1991), suggesting that the immune signaling mechanisms of these pathways have been conserved across species. Searching for the *Toll* gene sequence in the human genome, Medzhitov identified the human homolog of *Toll* (Medzhitov et al., 1997). hToll and dToll both transcribe a transmembrane protein with an extracellular N-terminal leucine-rich repeat (LRR) domain and an intracellular Toll/IL-1 receptor (TIR) domain. Following the discovery of the human homolog of *Toll* many members of the Toll-like Receptor (TLR) family were identified (10 in human, 12 in mouse) (Beutler, 2004; Takeda et al., 2003). A mouse strain that was not responsive to LPS was later found to have a mutation in its *Tlr4* gene (Poltorak et al., 1998). Work done by Akira and colleagues later showed that the TLR4 receptor specifically binds LPS (Akira et al., 2001). The discovery by Akira marked the first characterized PAMP-PRR combination. Many other receptor-ligand combinations followed (reviewed in Akira, Uematsu, & Takeuchi, 2006) . The TLR4-LPS interaction, consequently, became the protype receptor-ligand interaction in innate immunity and continues to be the most studied pathway in innate immune activation.

Structure and binding of LPS and TLR4

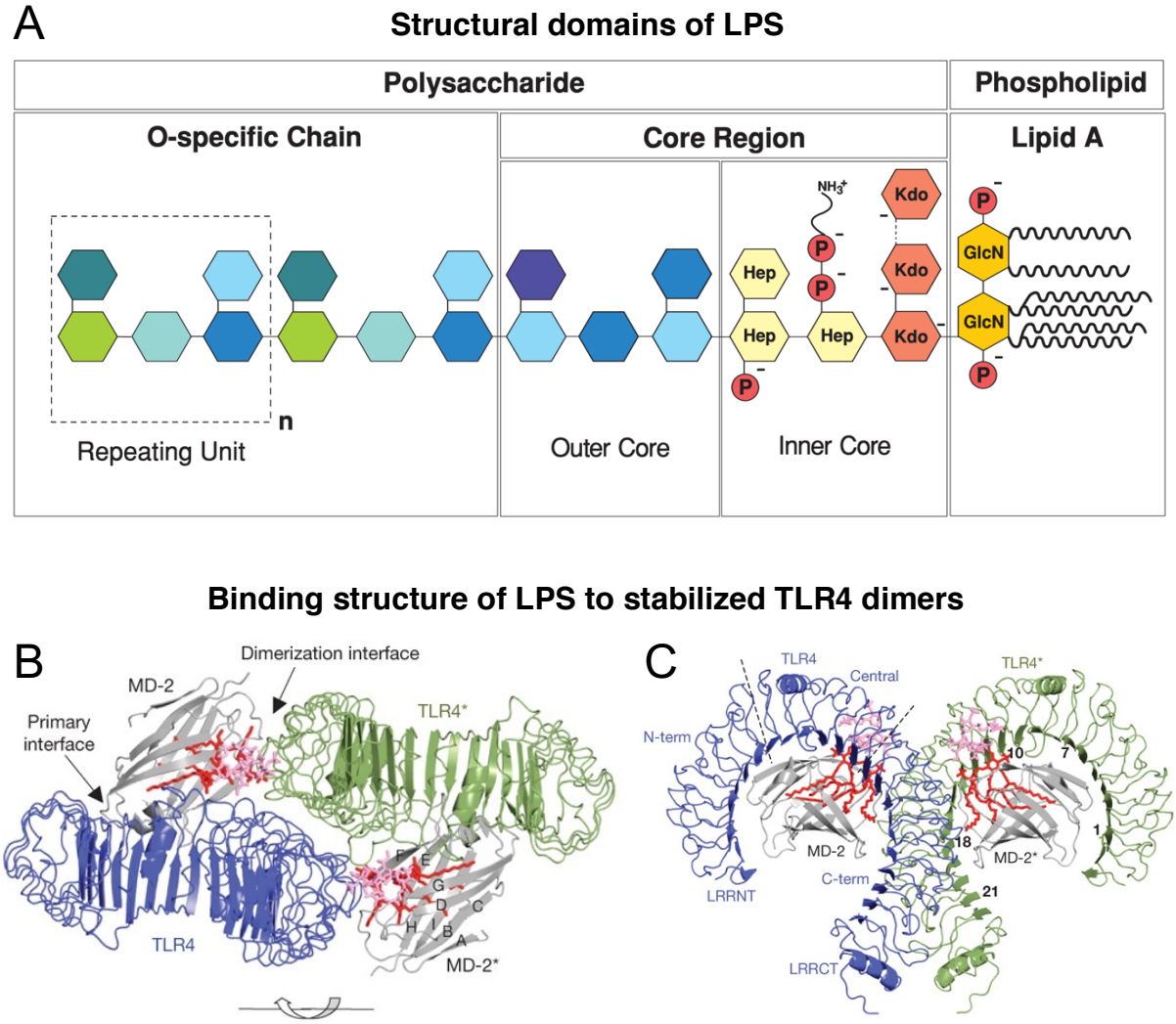


Figure 1.1 LPS domains and its binding to TLR4 dimers stabilized by MD-2
Adapted from Alexander 2001 and Park 2009.

(A) The chemical structure of LPS and its domains. The immunostimulatory membrane anchored Lipid A domain includes covalently linked poly and oligosaccharides bound to glucosamine (GlcN) residues. The core domain is bound to the lipid domain by three Kdo and D-glycero-D-*manno*-heptose (Hep) sugars and separated by an inner and outer core. The core domain is connected to the O-specific chain which is variable and comprised of up to 50 repeating motifs. (B) Top view of the stably dimerized structure of TLR4, MD-2 following binding of LPS. TLR4 binds MD-2 at the primary interface. Binding to LPS (red) leads the TLR4-MD-2 heterodimer to dimerize with another TLR4-MD—2 structure by binding at the dimerization interface. (C) The "horseshoe" configuration of the stabilized TLR4 receptor following binding to MD-2 and LPS recognition. This configuration creates hydrophobic regions for the phosphate groups and lipid chains of LPS Lipid A domain to bind.

1.3 TLR4 Pathway: signaling and regulation

1.3.1 MyD88/Mal vs. TRIF/TRAM dependent signaling

While TLR4 directly recognizes LPS, overexpression of TLR4 alone in embryonic kidney 293 cells did not increase responsiveness to LPS (Akira et al., 2001) suggesting that a series of adaptors and downstream effectors are required to activate the TLR4 pathway. Accessory proteins CD14 and LPS Binding Protein (LBP) are required to bring LPS to the receptor (Fenton & Golenbock, 1998; Ulevitch, 1993; Wright et al., 1990). Prior to binding with LPS, TLR4 forms a complex with myeloid differentiation factor 2 (MD-2) which creates the hydrophobic pocket where LPS can bind (Figure 1.1B and C) (Park et al., 2009; Shimazu et al., 1999). Following the binding of LPS, TLR4-MD2-LPS stably dimerizes with another TLR4-MD2-LPS structure via binding of the two intracellular TIR domain regions of TLR4 (Xu et al., 2000; H. Zhang et al., 2002). The dimerized version of TLR4, which spans across the cell or endosomal membrane, can then engage with macromolecular structures in the cytoplasm that have a binding affinity for the TIR domains of TLR4 (Bryant et al., 2015; Núñez Miguel et al., 2007; Song & Lee, 2012).

On the cytosolic side of the dimerized TLR4 proteins, a cascade of signaling events occur to initiate the pro-inflammatory activity of the cell. The transcription factor nuclear factor kappa B (NFκB) was independently shown to be activated by LPS (Patrick A. Baeuerle & Baltimore, 1991). NFκB was also recognized early as a downstream effector of TLRs given its signaling similarity to the function of *dorsal*, a transcription factor acting downstream of *Toll* in *Drosophila* (Gay & Keith, 1991; Ghosh et al., 1990). How the activated form of TLR4 culminates in the activation of NFκB has been the subject of extensive research.

There are currently two known paths through which activated TLR4, depending on its localization, initiates downstream signaling inside the cell. TLR4 anchored in the cell membrane and activated by extracytosolic LPS engages through its TIR domain with myeloid

differentiation primary response protein 88 (MyD88) and its adaptor protein MyD88-adaptor-like (MAL) (Fitzgerald et al., 2001; Medzhitov et al., 1998; Muzio et al., 1998). Following the binding with MAL, MyD88 recruits IL-1R-associated kinase (IRAK) family members through death domain (DD) interactions (Gottipati et al., 2008). This super-molecule does not bind in a 1:1 ratio, increasing evidence shows that these components come together to form a multiunit "myddosome" through which it engages its downstream effectors (Bryant et al., 2015; Gay et al., 2011). Following activation of the myddosome, TLR4 is internalized into endosomal vesicles, whereby the TIR domains of stably dimerized TLR4 bind with TIR domain-containing adapter inducing IFN-β (TRIF) protein and TRIF-related adaptor molecule (TRAM) (K. A. Fitzgerald et al., 2003; Kagan et al., 2008; Oshiumi et al., 2003; Yamamoto et al., 2002). These two paths for initiating the cellular response of activated TLR4 are respectively known as the MyD88-dependent and MyD88-independent pathways.

The divergent TLR4 response pathways engage different effectors and lead to a diversity of responses to LPS sensing. MyD88 and MAL signal though IRAK family proteins which activate the TNF-receptor associated factor 6 (TRAF6) (Keating et al., 2007; A. Liu et al., 2012; Verstak et al., 2014). TRAF6 acts as an E3-ubiquitin ligase which through a series of lysine-63 (K63) ubiquitin chains and the activation of mitogen-activated protein kinases (MAPKs) recruits the TGFβ-activated kinase 1 (TAK1), TAK1-binding protein 2 (TAB2) and TAB3 (TAK1/TAB2/3) complex. The TAK1/TAB2/3 complex then activates the IκB kinase (IKK) complex (Deng et al., 2000; Ostuni et al., 2010) which through lysine-48 (K48) ubiquitination targets IκB for degradation by the proteasome (Whiteside et al., 1997). IκB is attached to NFκB and is what keeps it, in the absence of TLR4 activation, sequestered in the cytoplasm. Following the degradation of IκB, NFκB rapidly translocates to the nucleus where it binds DNA in its function as a transcription factor and mediates the expression of pro-inflammatory genes (P. A. Baeuerle, 1995; Patrick A. Baeuerle & Baltimore, 1991; Ghosh et

al., 1998). The activation of TRAF6 and binding with TAK1/TAB2 also activates MAP3K which recruits the MAPK signaling cascade (Arthur & Ley, 2013; Takaesu et al., 2000). Engagement of the MAPK signaling cascade activates the MAPKs JNK and p38 (Sato et al., 2005). JNK and p38 activate the transcription factors activator protein 1 (AP-1) and cAMP response element-binding protein (CREB), respectively, which further mediates expression of pro-inflammatory genes (Das et al., 2009; Kim et al., 2008). The TRIF-TRAM pathway (i.e. MyD88 independent pathway) activates TRAF3 which, through IκB kinase-ε (IKKε) and TANK-binding kinase-1 (TBK1), activates the transcription factor interferon regulatory factor 3 (IRF3) which translocates to the nucleus (Katherine A. Fitzgerald et al., 2003). TRIF also activates the phosphatidylinositol 3-kinase (PI3K) which phosphorylates Akt which then modulates the expression of pro-inflammatory cytokines (Martin et al., 2003) (see section 1.3.2 on negative regulation of TLR4 signaling). These different pathways activated downstream of TLR4 activation by LPS also show crosstalk (Cheng et al., 2015; Dorrington & Fraser, 2019) and downstream signal integration (Sakai et al., 2017). Together these pathways activate a constellation of transcription factors that initiate the full scale of pro-inflammatory transcription in response to LPS.

1.3.2 Negative regulation of TLR4 signaling

The cytoplasmic sequestration of NFκB by IκB, and the subsequent release and activation of NFκB upon stimulus-dependent IκB degradation, has served as a prototypic model for how the TLR4 response pathway is negatively regulated against constitutive activation. Additional proteins that function as inactivators of inflammatory effectors that are then removed by degradation in response to sufficient stimuli of TLRs have been identified (Liew et al., 2005). Toll-interacting protein (TOLLIP) suppresses the kinase activity of IRAK and following LPS treatment TOLLIP is phosphorylated by IRAK1 and targeted for degradation, thus releasing

IRAK to complex with MyD88 (G. Zhang & Ghosh, 2002). Conversely, proteasomal degradation of positive effectors of TLR4 signaling in the absence of stimulus has also been observed. In the absence of robust TLR4 activation, the zing-finger protein A20 depresses TLR4 signaling by targeting for degradation the Ubiquitin-conjugating enzyme Ubc13 which is critical for building the ubiquitin chain of TRAF6 so that it can engage with the TAK1/TAB2/3 complex (Shembade et al., 2010; Wertz et al., 2004). In an interesting reversal of how many discoveries in TLR signaling arose from homology and comparison with *Drosophila*, the discovery of regulation by A20 in vertebrates led to the discovery of a similar model of regulation in *Drosophila* signaling by *Toll/Imd* (Chen et al., 2017). Alternative modes of negative regulation have also been identified such as the translocation of TLR4 to the lysosome via Rab7b where the receptor is assigned for degradation (Y. Wang et al., 2007), negative feedback of TLR4 activity via phosphoinositide 3-kinase (PI3K) (Laird et al., 2009), and deubiquitination of TRAF6 by the deubiquitinase cylindromatosis protein (CYLD) (Kovalenko et al., 2003) (in the presence of LPS stimulation CYLD is targeted for proteasomal degradation via Caspase 8 (O'Donnell et al., 2011)). These factors emphasize the interplay between positive effectors of the TLR4 response and negative regulators at different junctures in its signaling, and the central importance of signal induced protein degradation as a means of regulation.

1.4 Transcription and alternative splicing in macrophages

1.4.1 Escalation of core cellular processes define and drive the macrophage inflammatory state

Transcriptional induction of the inflammatory cytokine TNFα was only the first of an extensive panel of genes that LPS treatment was found to induce (B. Beutler et al., 1985). Simultaneously, many additional transcription factors beyond NFκB have been found to be

activated in response to TLR4 engagement (Guha & Mackman, 2001; Sweet & Hume, 1996). The genes whose expression is effected by LPS have been extended from the hundreds initially identified (Nau et al., 2002), to over a thousand genes (Bjorkbacka et al., 2004). While a robust transcriptional response is to be expected in the activation of an altered cell state, Liu and colleagues found that the scale of macrophage gene expression change in response to LPS+IFNγ was 1-2 orders of magnitude greater than other treatments and cell fate inductions in the same cell type (H. Liu et al., 2018) (Figure 1.2). In parallel to these findings, work has been emerging to elucidate the mechanisms of how the macrophage shifts its metabolic program to meet the energy needs of this large scale transcription induction program (Kelly & O'Neill, 2015). In addition to activating a switch to glycolytic metabolism (Krawczyk et al., 2010), recent findings have suggested that LPS treatment also directly engages and increases the production of specific pro-glycolytic metabolites (Infantino et al., 2011; Michelucci et al., 2013; Tannahill et al., 2013). These findings emphasize the large-scale signaling shift in many of the core cellular processes activated by TLR4 during and inflammatory response.

A key cellular process not discussed so far, which is relevant in the context of robust transcriptional induction, is alternative splicing. Alternative splicing is a process of exon excision and inclusion of mRNA by the spliceosome and controlled by *cis* and *trans* acting proteins on the mRNA (Black, 2003). Estimates of human protein coding genes that undergo alternative splicing range from 60% (Modrek & Lee, 2002), to upwards of 90% when looking at multi-exonic genes that have two or more known isoforms (Lee & Rio, 2015). At the proteomic level it is estimated that alternative splicing expands the number of proteins that can be generated from the human genome by a factor of 10 (Nilsen & Graveley, 2010). It would be expected that in macrophage transitions, alternative splicing events could play a critical role in the determination of fate and activation states. A surprising finding, however, came from Lin and colleagues who did a transcriptome-wide analysis of alternative splicing in macrophage polarization and found that alternative splicing was highly and preferentially deployed in M1 activation (pro-inflammatory, the one associated with the LPS treatment) and only minimally so in M2 macrophage determination (Lin et al., 2016). The escalated deployment of alternative splicing in the activation of inflammation was similarly observed by

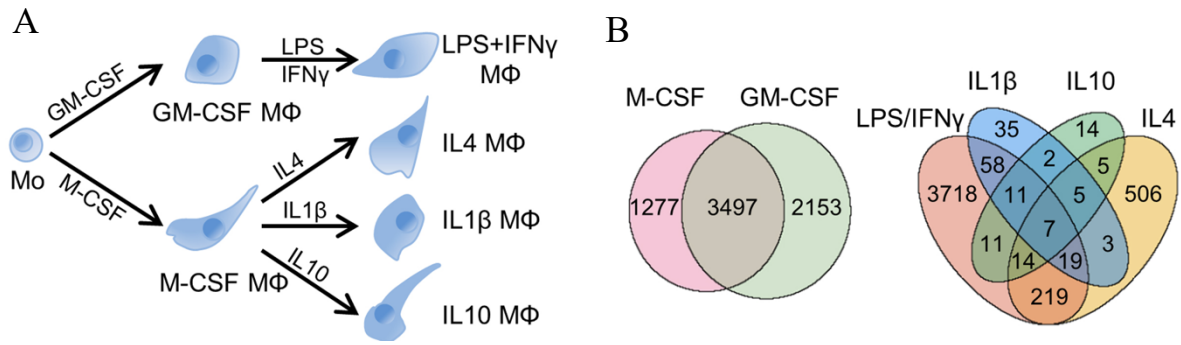


Figure 1.2: **Elevated transcription in LPS treated macrophages.**
From H. Liu et al., 2018

(A) Diagram of differentiation and polarization of human primary monocytes (Mo) to macrophages (Mϕ) in culture. (B) Venn diagrams with numbers of differentially expressed genes with fold change ≥2 (q-value ≤0.05) for GM-CSF and M-CSF treated macrophages versus monocytes (left), and for the four polarizations (right).

Liu and colleagues who found that in macrophages treated with LPS+IFNγ, alternative splicing events exceeded by an order of magnitude the number of events observed following other cell fate and activation treatments of monocytes (H. Liu et al., 2018) (Table 1.1). These findings suggest that, in addition to the known core cellular processes (proteasomal degradation, elevated transcription, and glycolytic metabolism) used by macrophages in their rapid transition from homeostatic surveyors to activators of inflammation, alternative splicing is likely to also play a key role.

Cell Type	Culture	Treatment	Exon Inclusion events	Exon Skipping events
Human primary monocytes	Mo	GM-CSF	2574	1794
	Mo	M-CSF	2264	1433
	GM-CSF	LPS/IFNγ	448	357
	M-CSF	IL4	34	27
	M-CSF	IL1β	25	13
	M-CSF	IL10	17	12
THP-1 cell line	THP-1	PMA3D	305	267
	THP-1	PMA5Dr	349	211
	THP-1	VD3	1142	519
	PMA3d	LPS/IFNγ	224	131
	PMA3d	IL4/IL13	133	75

Table 1.1 **Comparison of differential alternative splicing events in monocyte (Mo) and macrophage treatments.**
From H. Liu et al., 2018

Liu and colleagues compared the number of alternative splicing events observed in monocyte derived and immortalized macrophages under various differentiation and stimulus conditions. While the differentiation transitions (GM-CSF and M-CSF) led to the highest number of alternative splicing events in monocyte derived macrophages, among the stimulus treatments LPS/IFNγ led to the highest number of alternative splicing events.

1.4.2 Dynamic regulation of signaling by alternative splicing, in macrophages and other contexts

Alternative splicing has been previously shown to have a regulatory role in TLR4 signaling through variants of critical effectors such as MyD88, IRAK1, MD2, and TLR4 itself. MyD88s, an alternatively spliced variant of MyD88 that lacks its critical intermediate domain, binds IRAK1 but fails to induce IRAK phosphorylation, thus acting as a negative inhibitor of MyD88-dependent signaling (Burns et al., 2003; Janssens et al., 2002). As compared to healthy donors, patients with acute respiratory distress syndrome (ARDS) have been shown to have higher ratios of MyD88l -the complete and functional version- to the splice variant MyD88s (Blumhagen et al., 2017). IRAK1 was also found to have a splice variant that has a dominant negative effect on TLR4 signaling. IRAK1c lacks exon 11 of the IRAK1 gene and fails to induce phosphorylation, thus truncating TLR4 signaling at the myddosome formation stage (Rao et al., 2005). A similarly truncated variant that negatively impacts TLR4 signaling was described for the TLR4 accessory protein MD-2. The alternatively spliced MD-2s variant lacks exon 2 of the MD-2 gene and fails to induce a stable ligand binding pocket for LPS when complexing with TLR4 (Gray et al., 2010). A splice variant of TLR4 itself has also been identified. Soluble mTLR (smTLR4) includes an additional exon that adds a stop codon after exon 2 and inhibits activation by LPS. smTLR4 was also found to be expressed as a result of LPS stimulation suggesting that this might be used as way to terminate TLR4 signaling during the inflammatory resolution phase (Iwami et al., 2000; Jarešová et al., 2007). Patients with rheumatoid arthritis were found to have several splice variants of the A20 gene (Yoon et al., 2013), though it remains to be seen whether alternative splicing of A20 is also a regulatory mechanism of TLR4 signaling in healthy patients.

The TLR4 pathway related examples listed above all involve switching between a functional variant and alternatives that lack comparative function and stability. In some cellular

contexts, alternative splicing was observed to have a regulatory effect by switching to attenuated variants. In other cellular contexts, however, alternative splicing has been shown to have regulatory capacity that goes beyond switching between functional and non-functional variants. Alternation in splicing can also produce variants with different positive signaling results. The inhibitor of apoptosis protein (IAP) survivin was shown to have two variants that are both functional, but differ in their localization within the cell, suggesting that alternative splicing can be used to direct a protein to different localizations in different contexts (Mahotka et al., 2002). It remains to be discovered whether macrophage activation is also regulated by alternative splicing in such a manner.

1.4.3 The spliceosome network of proteins drives specificity in different contexts

Alternative splicing events are understood to be guided by transcript sequences and RNA binding proteins that differentially promote the inclusion or exclusion of exons and introns. There are *cis*-regulatory sequences known as exonic splicing enhancers (ESE) and intronic splicing enhancers (ISE) which promote exon and intron inclusion, respectively. Exonic splicing silencer (ESS) and intronic splicing silencer (ISS) are *cis-regulatory* elements that promote exon and intron repression, respectively (Black, 2003; Yabas et al., 2015). There also trans-acting factors which include the serine-arginine (SR) rich proteins and heterogeneous ribonucleoprotein particle (hnRNP) proteins which recruit the spliceosome to different splice sites to promote, in the case of SR proteins, splice site usage and, in the case of hnRNPs, splice site skipping (Mayeda et al., 1999; Jun Zhu et al., 2001) (Figure 1.3). All of these factors are targets for regulation of signaling in innate immunity (Fang et al., 2017; Grohar et al., 2016; Yabas et al., 2015). The spliceosome complex itself has recently be shown to also be able to dynamically regulate splicing and for its regulatory impact to not simply be relegated to whether it is or isn't recruited to specific sites by ancillary regulators. Work by Papasaikas and

colleagues found that the spliceosome, far from being a rigid and universal structure, consists of a network of persistent and transient proteins that are engaged for complete or just parts of the splicing reaction, respectively. More surprisingly, silencing by siRNA of persistent or transient components affected alternative splicing in qualitatively different ways (Papasaikas et al., 2015). These findings suggest that the selective engagement of the core spliceosome proteins also plays a role in regulation by alternative splicing. This conclusion is supported by an earlier finding by De Arras and colleagues that inhibition of *Sf3a1* and *Sf3a2*, two genes transcribing proteins essential for spliceosome function, diminished IL-6 production in LPS treated macrophages (De Arras et al., 2013). More recent work by Liu and colleagues found that muscleblind like splicing regulator 1 (MBNL1) is specifically essential for the monocyte to macrophage transition (H. Liu et al., 2018). The independent findings by Papasaikas, De Arras, and Liu emphasize the role core splicing factors can play in selective regulation in different contexts.

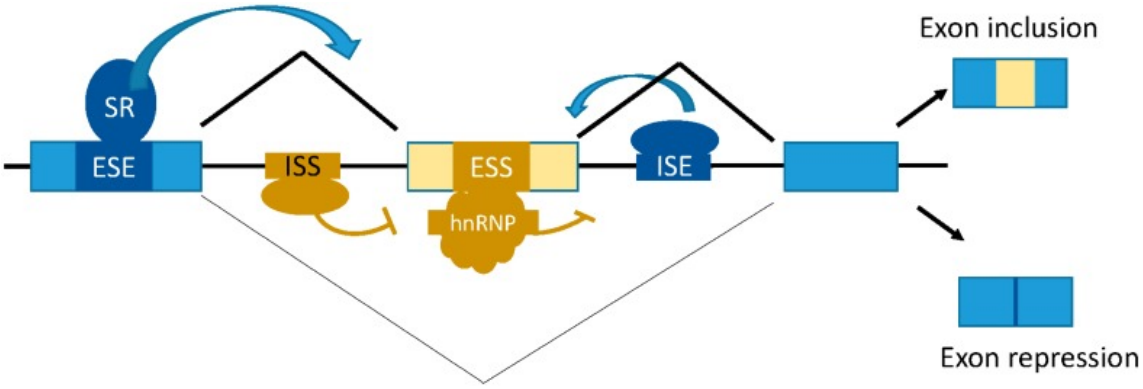


Figure 1.3: ***Cis* and *Trans* regulatory mechanisms for alternative splicing.**
From Yabas et al., 2015

Mechanism of alternative splicing of pre-mRNAs. The *cis*-regulatory elements that control alternative splicing are composed of unique nucleotide sequences. The exonic splicing enhancers (ESE) and intronic splicing enhancers (ISE) (dark blue) respectively promote exon inclusion (light blue boxes), while exon repression requires the exonic splicing silencer (ESS) and intronic splicing silencer (ISS) elements (brown). The *trans*-acting factors that promote exon inclusion are the Serine (S)-Arginine (R) SR rich proteins while hnRNP proteins can promote exon skipping. In the diagram the middle exon can be included to give rise to a polypeptide with 3 exon units, while the alternatively spiced polypeptide skips the variable exon to yield a protein with 2 exon units.

1.4.4 Splicing aberrations and inflammatory dysregulation in myelodysplastic syndromes

Beyond the *in vitro* findings mentioned in the previous section, mutations in core spliceosome factors have also been shown to play a key role in the pathogenesis of disease. Myelodysplastic syndromes (MDS) are a group of diseases whose pathogenesis is associated with altered proliferation and differentiation of bone marrow cells. Though the pathogenesis and disease phenotype is quite diverse across patients, it often involves cytokine and immune system abnormalities, ring sideroblasts, anemia, and bone marrow dysplasia (Tefferi & Vardiman, 2009). Early work has shown that anemia and ring sideroblast – two of the pathogenic categories of MDS – is often associated with patients having a mutation in the Splicing Factor 3b Subunit 1 (SF3B1) gene (Visconte et al., 2012). Further studies found that mutations in SF3B1 are often associated with MDS (Arber et al., 2016). Two additional splicing related genes, U2 Small Nuclear RNA Auxiliary Factor 1 (U2AF1) and Serine And Arginine Rich Splicing Factor 2 (SRSF2), were also found to have mutations in subset of MDS patients (Papaemmanuil et al., 2013).

Moreover, recent studies have begun to show dysregulated TLR signaling in patients with MDS. One study found that downstream signalers of TLR4 are overexpressed in patients with MDS (Velegraki et al., 2013), another study found broad dysregulation of TRAF6 – a critical signal mediator in TLR signaling- in patients with MDS (Culver-Cochran & Starczynowski, 2018). These parallel findings suggest that the pathogenesis of MDS may be correlated with the dysregulation of TLR4 signaling by aberrant alternative splicing, though the mechanism by which these mutations lead to this dysregulation, and how it deviates from normal TLR4 signaling, remains to be elucidated.

1.5 Genome-wide perturbation studies of the TLR4 response to LPS

1.5.1 Categorizing the canonical TLR4 pathway

As a corollary to the outlined incremental insights so far to the canonical TLR4 signaling pathway, various schema have been proposed to integrate the myriad of effectors, mediators, and regulators shown so far to bridge LPS recognition with pro-inflammatory transcription. The Kyoto Encyclopedia for Genes and Genomes (KEGG) curates a TLR signaling pathway that includes 104 members (KEGG) with only the essential effectors downstream of TLR4 included (Figure 1.4A). Sun and colleagues took a more methodical approach by curating known and putative effectors in TLR signaling, and running them through a targeted RNA interference (RNAi) screen in human and mouse macrophages treated with a panel of TLR ligands (J. Sun et al., 2016). The canonical TLR pathway that emerges from their study includes 126 genes, with some being shared and many being specific for different TLR ligands. Within the same ligand treatment, some effectors were also shown to be species-specific in their requirement (Figure 1.4A). Sun and colleagues propose a model of TLR signaling that includes the spectrum of TLR signaling factors segmented into three modules (Figure 1.5A-B), the encoder module which includes the TLR receptors and their proximal signaling adaptors (such as MyD88/MAL and TRIF/TRAM), the transmission module (including factors such as NFκB, MAPK and PI3K/Akt), followed by the decoder module (which includes the transcription factors and cytokines). Working in dendritic cells (DC), Mertins and colleagues aimed to do a similar analysis by curating known TLR signaling components and looking for "orphaned" effectors (Mertins et al., 2017). (Orphaned in the sense that their matching kinases haven't been identified.) Combining ss (SILAC)-based phosphoproteomics and computational network analysis, they compiled a list of 161 genes associated with LPS responsiveness. Breaking down the network into three groups; "seeds", "intermediates", and "transcriptional regulators" (Figure 1.5C-D), these groups map similarly to the "encoder", "transmission" and "decoder"

model that Sun et al. proposed. These similar models summarize the understanding of innate immune activation that has emerged from the accrued findings so far, that it comprises a recognition-to-signaling-to-transcription system. These two studies also highlight different blindspots in the currently curated models. The weighted-network work by Mertins highlights how the itemized identification of different factors has left gaps in our understanding of how many of these effectors interact and how these factors coalesce into a functional signaling network. The comparative work by Sun, on the other hand, highlights how insights gained from dispersed species and cell types may come up short when extrapolated into a cohesive vertebrate TLR signaling pathway. To elucidate a more complete model of innate immune activation by TLR4 signaling, both of these considerations must be taken into account.

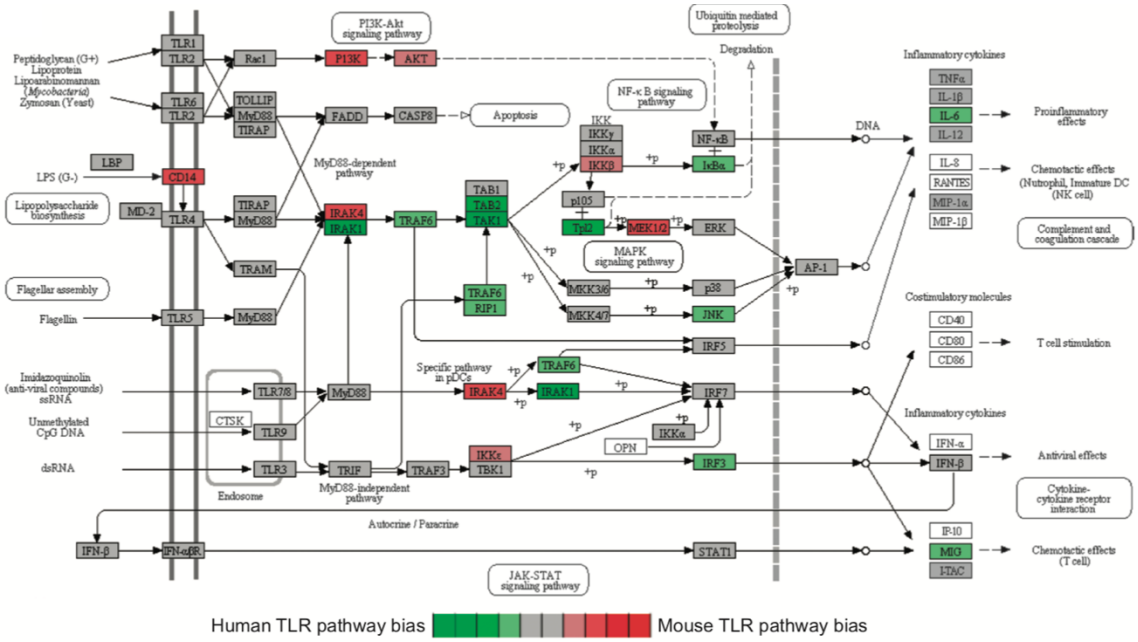


Figure 1.4: **Canonical TLR signaling pathway in human and mouse.**
From KEGG, Sun et al., 2016.

The canonical TLR pathway from the KEGG pathway database, overlaid with species specific dependency found by Sun et al., 2016.

Network models of the TLR signaling pathway

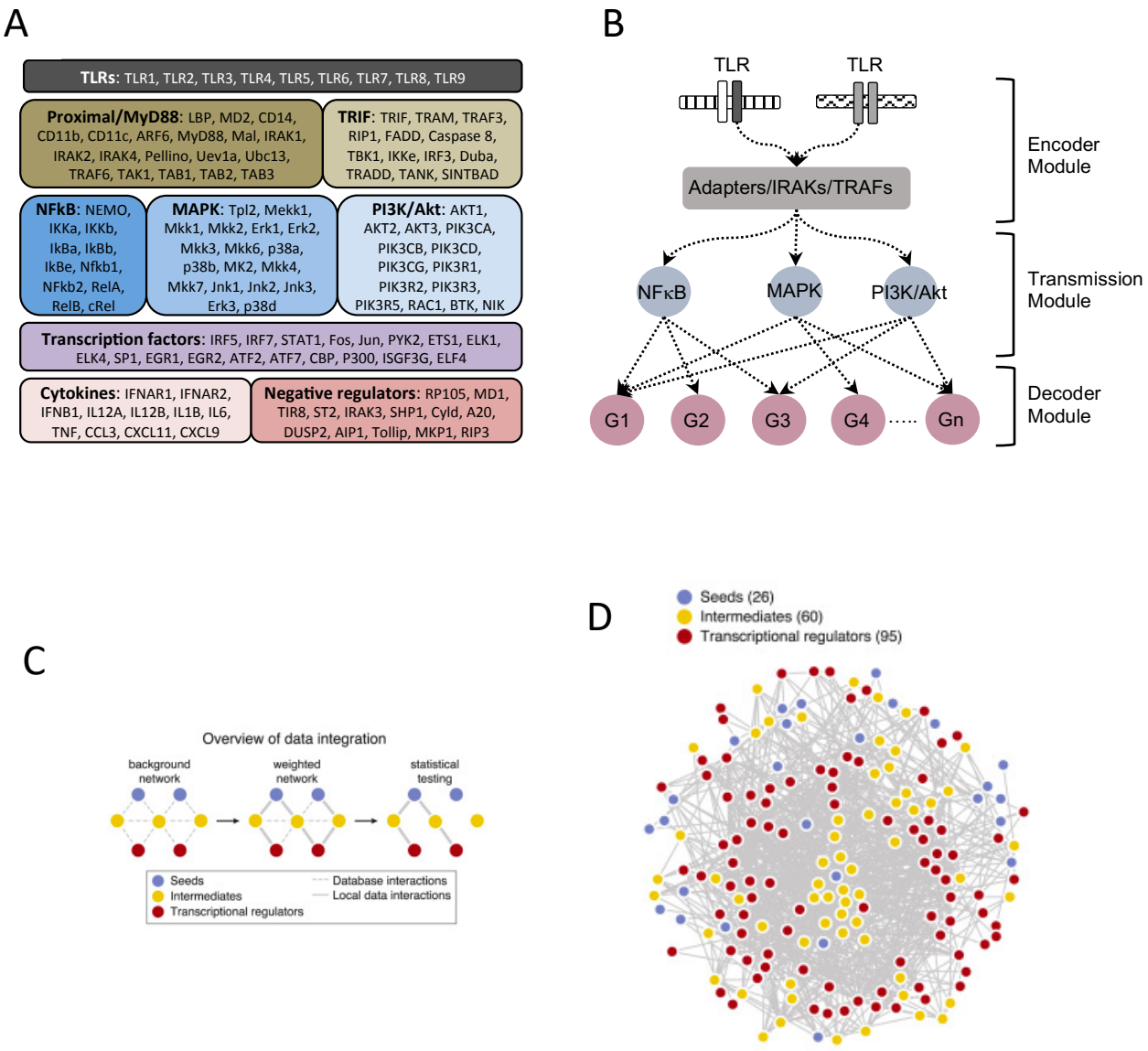


Figure 1.5: **Models of the canonical TLR signaling pathway.**
From Sun et al., 2016, Mertins et al., 2017

(A) Schematic by Sun et al., 2016 of signal flow in the TLR pathway (B) 126 canonical TLR signaling genes curated by Sun et al., 2016. (C) The computational framework for integrative analysis of the functional and physical proteomics applied by Mertins et al., 2017 for analysis of canonical TLR signaling (D) A TLR signaling transduction network proposed by Mertins et al., 2017 that connects 27 seeds (blue) to 95 transcriptional regulators (red) through the top 60 intermediate (yellow) nodes.

1.5.2 High-throughput assays to identify novel regulators in immune cells

The regulatory insights and discoveries outlined so far have all contributed to the incremental reconstruction of the canonical TLR4 signaling pathway. A necessary complement to that approach, however, has been the use of unbiased genome-scale studies. RNA interference (RNAi) and clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 assays make it possible to target the expression of specific genes in the genome (Cong et al., 2013; Fire et al., 1998; Jinek et al., 2012). Libraries of short interfering RNA (siRNA) or guide RNAs that cover the complete (or near complete) human or mouse genome make it further possible to expand these technologies to high-throughput assays and identify novel candidates in a comprehensive and unbiased manner (Berns et al., 2004; Boutros et al., 2004; Shalem et al., 2014; T. Wang et al., 2014). In attempting to expand our insight into the regulatory mechanisms of TLR4 signaling, it is critical to fully utilize these genome-level approaches so that novel regulators of the pathway can be discovered and signaling gaps can be filled in.

Adapting these gene perturbation systems to macrophages cells, however, faces a critical obstacle. Both platforms rely on exogenous RNA delivery to guide the perturbation mechanism to their target (Shalem et al., 2014; Silva et al., 2004). This presents a challenge in the context of studying immune cells, specifically in the context of TLR signaling, as some TLRs - as a form of host defense (especially against viruses)- bind and respond to the presence of exogenous RNA (Alexopoulou et al., 2001; Heil et al., 2004). To circumvent this challenge, Li and colleagues used a systematic approach to identify an siRNA delivery system that avoids activating the cells' intrinsic immune response (N. Li et al., 2015). Using that system, Li and colleagues also developed two reporter cell lines in human and mouse macrophage-like cells that can be used for high-throughput assays of TLR4 activation. In the mouse cell line, they introduced two reporters associated with downstream effectors of TLR4. A green fluorescent protein (GFP) fused to the relA NF-κB transcription factor driven by its endogenous promoter

and a secondary reporter cassette that includes the TNFα promoter driving expression of the single exon red fluorescent protein mCherry. The first reporter enables by high content imaging the tracking of NF-κB cytosol-to-nuclear translocation, while the second reporter enables the tracking of initiated TNFα transcription through the red fluorescence channel. In human cell lines, Li and colleagues used a dual luciferase reporter system, including a firefly luciferase driven by the human TNFα promoter and a renilla luciferase driven by a ubiquitin promoter. The former is used as a readout for TNF transcription and the latter is used as a normalization control (Figure 1.6). It should be noted that treatment with LPS did not lead to increased activation of ubiquitin driven renilla luciferase, enabling its use as a normalization control (Figure 1.6G). Renilla activation showed a slight decrease in some wells treated with LPS compared to untreated, though that may be driven by the lower cell count in some of the LPS treated wells. Utilizing these platforms, Li and colleagues (Ning Li et al., 2017) and Sun and colleagues (J. Sun et al., 2017) completed genome-wide analysis of the LPS induced response in mouse and human macrophages, respectively (Figure 1.7).

Using a different cell and perturbation system, Parnas and colleagues published a genome-wide CRISPR study in mouse dendritic cells treated with LPS (Parnas et al., 2015). Critical differences exist in the design and assay of the Parnas study and the three LPS studies done in macrophages by Li and Sun. Parnas and colleagues used bone marrow derived cells from mice that were then differentiated into dendritic cells as opposed to macrophages, and used a CRISPR library instead of the RNAi platforms used by Li and Sun. These four studies, collectively, present the first chance to study the TLR4 signaling pathway in a comprehensive way and expand our insights into its regulatory program (see chapter 3 and chapter 6).

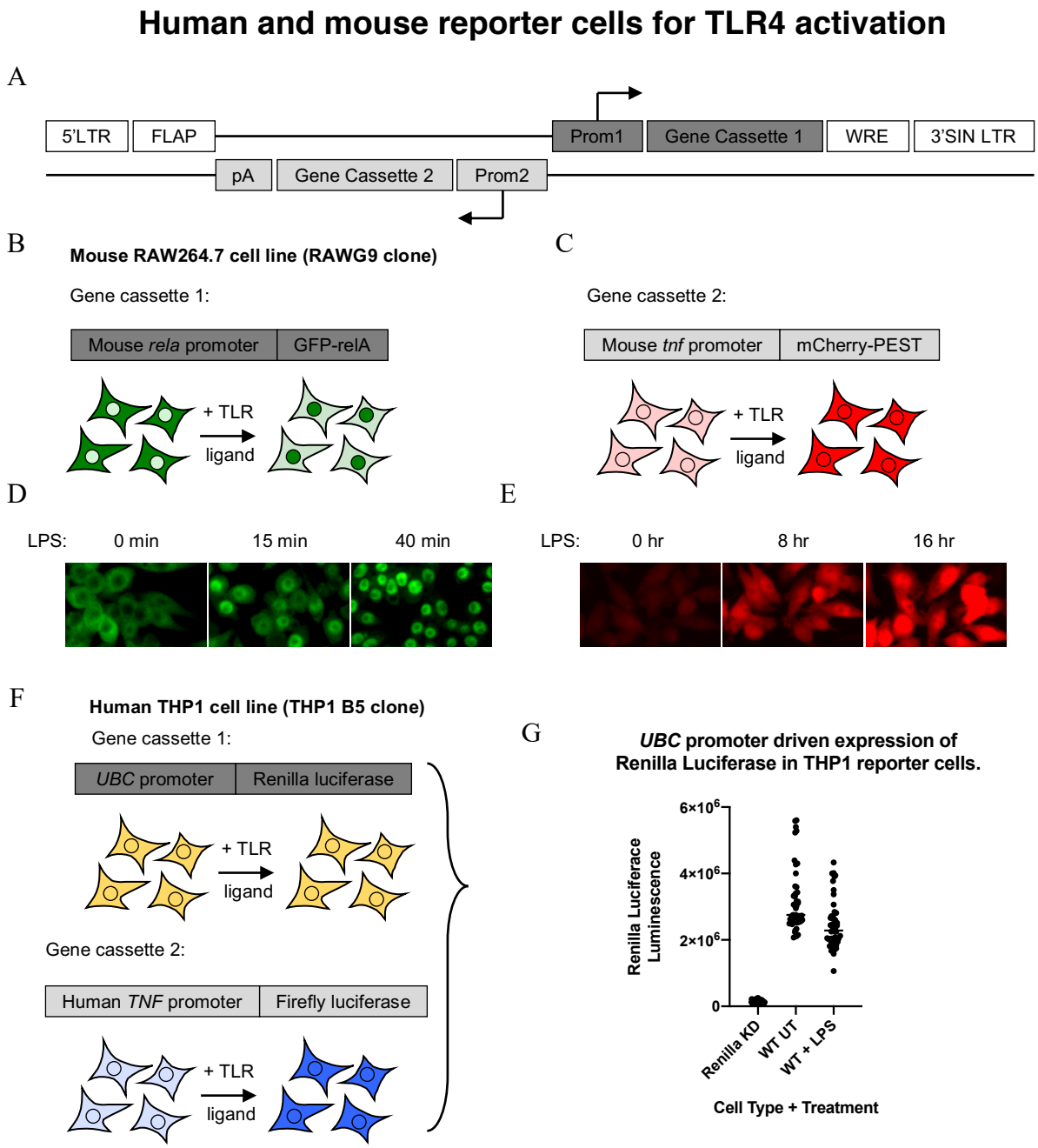


Figure 1.6: **Reporter cell lines for tracking TLR4 activation in human and mouse macrophages.**
From Ning Li et al., 2015

(A) Design of a dual-promoter lentiviral vector for expression of TLR pathway reporters. (B + C) Gene cassettes in the mouse RAW G9 reporter clone containing Bb) the mouse rela promoter driving expression of a GFP-relA fusion protein and (C) the mouse tnf promoter driving expression of an mCherry protein. (D) Cytosol-to-nuclear translocation of the GFP-relA fusion in RAW G9 cells up to 40 min after treatment with 10 ng/ml LPS (E) Increased tnf promoter-driven mCherry expression in RAW G9 cells up to 16 hr after treatment with 10 ng/ml LPS (F) Gene cassettes in the human THP1 B5 reporter clone containing the human UBC promoter driving constitutive expression of renilla luciferase and the human TNF promoter driving TLR ligand-inducible expression of firefly luciferase. (G) Luminescence of the UBC promoter driven Renilla luciferase in Renilla knock down THP1 cells and WT untreated and treated cells.

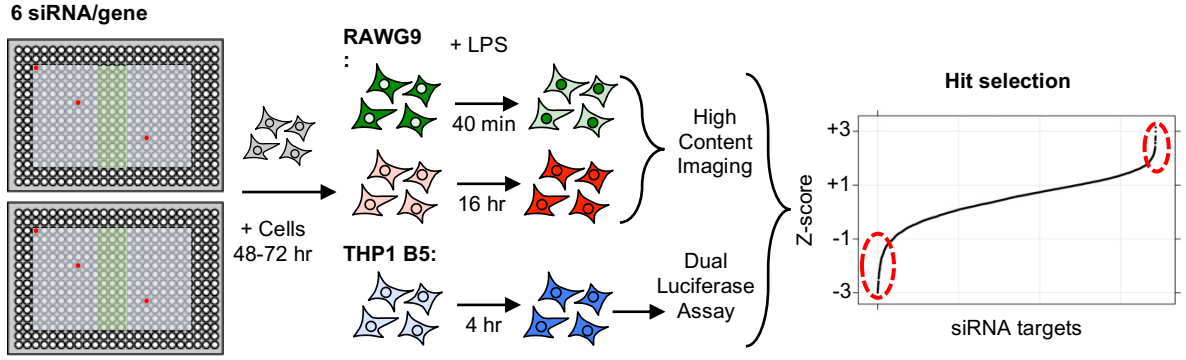


Figure 1.7: **A siRNA screening pipeline for human and mouse macrophage activation.** *From Sun et al., 2016*

Workflow for the RNAi screen using six siRNA sequences per gene distributed in separate regions of a 384-well plate. Reporter cell lines were assayed for their responses to simulation with LPS, and the effects of individual gene perturbations were measured

1.5.3 Finding regulators beyond the highest scoring and annotated hits of omic studies

As reported findings from the high-throughput studies of the response to LPS, the aforementioned Sun, Li and Parnas studies selected insights from the highest scoring hits and enrichments that their respective screens identified. The RNAi screen in macrophages led to the characterization of species-specific roles for different IRAK family proteins in human and mouse macrophages (J. Sun et al., 2016) and the genome-wide CRISPR study in mouse dendritic cells by Parnas implicates a role for components of the oligosaccharyltransferase (OST) complex in the LPS-induced expression of TNFα (Parnas et al., 2015). Beyond these specific findings, however, the vast scope of insights generated by these genome scale studies remain largely unexplored.

The gap between the genome-scale of measurements in high-throughput studies, and the modest number of reported candidates and insights that the studies culminate in, is not unique to the studies of the LPS response. Many hits from high throughput studies lack extensive annotation and also lack clear direction for how to contextually follow up on them

(Manly et al., 2004). Prioritization of candidates from high-throughput screens is further constrained by the high rates of false positives that these studies, especially using RNAi, have been associated with (Aguirre et al., 2016; A. L. Jackson & Linsley, 2004; Mohr et al., 2010). In prioritizing hits for further validation, researchers must find a balance between prioritizing candidates based on potential biological relevance and the feasibility to follow up on numerous gene candidates (Finkbeiner et al., 2015). These considerations contribute to a bottlenecking of screen data hit selection that is often most strongly guided by the *a posteriori* knowledge of the authors over the ranking and statistical results from the screen (Lotterhos et al., 2016). This phenomenon also explains what can be described as the "quantum leap" in reports of high-throughput studies, when the analysis leaps - with sparse analytical justification- from considering statistically prioritized lists of candidates to a handful of hits that are selected for further validation. While many insights and discoveries from high-throughput studies have been made through this approach, to fully realize the potential of these efforts to elucidate robust regulatory pathways, unbiased and comprehensive candidate prioritization approaches that address these challenges are necessary.

1.5.4 Score based hit selection: false positives, false negatives, and setting a cutoff

A basic approach towards unbiased hit prioritization from high-throughput studies relies on using a normalized scoring method that assigns a ranking to each candidate based on its score. Readout scores generated by high-throughput assays reflect a mix of biological variability and technical variation that contribute to each target's score. A commonly used corrective approach is the Z-score method (originally called Z-factor (J. H. Zhang et al., 1999)) which assigns a score to each target based on its score's deviation from the mean of scores for all the targets in the assay (Birmingham et al., 2009), thus correcting for the technical variability intrinsic to the assay. Variations on the Z-score approach have been developed to adjust for outliers and batch

and plate variability, such as robust Z-score (using median and median absolute deviation in place of the mean and standard deviation) and normalizing based on the score distribution of each plate as opposed to the complete assay (Birmingham et al., 2009; B. Dutta et al., 2016; Tseng et al., 2012). The challenge of the Z-score approach, however, is that it is dependent on the setting of a cutoff that defines the absolute score above which all candidates are considered hits and below which all candidates are considered non-hits (Boutros & Ahringer, 2008). The setting of a single cutoff is associated with an intrinsic compromise between decreasing the false positive rate while increasing the false negative rate or vice versa (Malo et al., 2006). On one hand, setting of a stringent cutoff for hit selection (such as the top 1% of hits or less) reduces the number of false positives amongst candidates but greatly increases the number of false negatives (i.e. targets assigned as non-hits that are biologically significant but may have a readout that is either less sensitive or obscured by the technical variation of the experimental design). Conversely, setting a more lenient cutoff goes hand-in-hand with the inverse error rate, the number of false negatives is reduced but the false positive rate is greatly increased. An additional impact of the more lenient cutoff strategy for hit selection is that it makes the follow up lower-throughput experiments less efficient, as a high number of candidate hits will fail to validate by secondary screening. The efficiency consideration drives many studies to rely on more stringent cutoffs and err on the side of a low false positive rate. This approach, however, leaves unexplored a large portion of potential candidates. When aiming to comprehensively characterize a signaling network it is critical to interrogate the lower scoring putative hits as well. Analysis by Rosenbluh and colleagues of genomic and proteomic studies to characterize the signaling network of β-catenin-active cancers found that many critical novel signaling components were in the lower scoring segments of the assay scores that are missed by a stringent cutoff (Rosenbluh et al., 2016). This finding emphasizes the need for an unbiased hit selection strategy that captures more of the lower scoring potential hits. Furthermore, the

artificial rigidity of a single cutoff, be it stringent or lenient, crucially obscures the more complex reality whereby targets identified by the screen exist on a spectrum of confidences with novel biological insights distributed across the range of assay scores. To more fully utilize the potential of high-throughput studies in characterizing novel regulatory mechanisms and candidates, an approach to hit prioritization that circumvents these challenges is needed.

1.6 Limited overlap in hits from comparative high-throughput studies

1.6.1 Overlap in reported hits of parallel studies is limited across fields

In considering the results from multiple genome-wide studies characterizing the TLR4 response to LPS, it is also critical to recognize that comparative analysis of high throughput studies in different systems have often had limited concordance. From microarray assays to gene-perturbation studies, analysis of published datasets from varying omics platforms repeatedly show modest overlap with limited statistical significance across parallel studies (Bhinder & Djaballah, 2013; Ein-Dor et al., 2006). High-throughput gene perturbation studies of similar biological phenomena that have surprisingly limited overlap have been found in studies of Influenza (Watanabe et al., 2010) and Human Immunodeficiency Virus (HIV) (Hirsch, 2010).

The unexpected heterogeneity of hits from related high-throughput studies has been broadly attributed to the varying experimental conditions of different research settings affecting the ranking of hits and thus obfuscating the true biological variability being measured (Ramasamy et al., 2008). Independent meta-analysis by Bushman and Hao of the Influenza and HIV studies found that the lack of overlap is largely driven by the false-negative rates in the hit selection methods of the comparative studies (Bushman et al., 2009; Hao et al., 2013). These insights further suggest that developing correctives for broader identification of hits beyond highest scoring targets is critical for robust and reproducible results from high-throughput

studies. It also suggests that an increase in the number of hits shared across parallel studies can be used as a metric for false negative correction.

1.6.2 Measuring significance of overlap by number of shared hits or by statistical enrichment

Robustly measuring the overlap of hits across two or more studies can be done in two ways. The number of shared hits can be counted and the result taken as a measure of the true-positive hits identified by the two studies (Rhodes et al., 2002). This approach is often referred to as "vote counting" when used to prioritize individual targets that appear as hits in more than one study (Ramasamy et al., 2008). While the approach of directly counting the number of shared hits does not address the number of false-positives each set of hits has and the false negatives the set excludes, it is still an important tool in assessing the sensitivity of high-throughput studies.

An alternative approach is a statistical test for how much the overlap across the two sets exceeds the null (i.e. the assumption that overlap between unrelated studies will be no greater than a random sampling of hits of equal size). A frequently deployed test for convergence and non-random overlap is the hypergeometric test, also known as the one sided Fisher's Exact Test (FET) (Birmingham et al., 2009; R.A. Fisher, 1925; T. Nguyen et al., 2015). The hypergeometric test measures the statistical significance of the overlap of two groups selected from two independent sets. In the case of comparing high-throughput studies the two independent sets would correspond to the two genome-scale studies and the two groups being compared correspond to the sets of hits selected from each group. This measurement relies on three critical numbers, m which is the number of shared members across the two sets, n_1 and n_2 which are the sizes of the two selected sets of hits, and n which is the group of candidates from which the set of hits was selected from (Fury et al., 2006) (Figure 1.8). Since the total

number of candidates measured in genome-scale studies are near similar across comparative studies, improvements in the statistical enrichment of agreement across parallel omic studies can be driven by either an increase in the number of shared hits (bigger m) or a decrease in the size of candidates selected as hits (smaller n_1 and n_2). Given the relative and complementary strengths of the two ways to measure shared enrichment – the number of shared hits which is driven by a decrease in false-negative rate and significance of overlap which is strongly driven by reduction in the false-positive rate - both measurements are critical proxies for assessing improved hit selection methods.

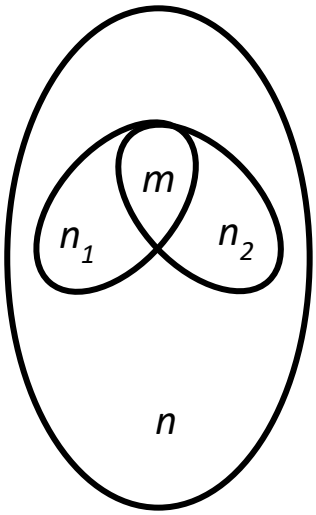


Figure 1.8: **Schematic of measurements calculated by the hypergeometric test of statistical enrichment.**

A schematic of the different components being measured when calculating the statistical significance by the hypergeometric test. n_1 and n_2 correspond to the two selected hit sets. m is the set of shared hits between the two selected sets. n is the set from which the hits are selected from in both groups.

1.6.3 Comparative studies of HIV dependency factors as example of hit selection challenge in high-throughput studies

An example often referred to in the literature of the high discordance of hit selection from high-throughput assays are a set of three independent studies looking at the essential proteins required for early infection of HIV, also known as HIV Dependency Factors (HDFs) (Hirsch, 2010; Jian Zhu et al., 2014). Published within months of each other, Brass *et al.* (Brass et al., 2008), König *et al.* (König et al., 2008), and Zhou *et al.* (Zhou et al., 2008) used genome-scale RNAi studies to identify host factors that were required by the virus for effective early stage infection of the host cell. The three siRNA studies for HDFs have been the subject of a number of subsequent meta-analysis studies (Bushman et al., 2009; Hirsch, 2010; Jian Zhu et al., 2014) and the three set comparisons serve as a validating example of the widespread challenge of the limited overlap found across related genome scale studies.

Since these studies have already been published in their entirety they can also be used as test data sets to evaluate how different hit selection approaches change the metric of overlap across screens. Various attempts have been made at developing benchmarking or synthetic datasets that can be used to evaluate the accuracy, sensitivity, and specificity of different hit selection approaches (Geistlinger et al., 2020; Mathur et al., 2018; T.-M. Nguyen et al., 2019; Roder et al., 2019). The setting of a gold standard dataset by which different prioritization methods can be compared remains one of the critical challenges in the bioinformatics of high-throughput data (Khatri et al., 2012; Mathur et al., 2018; Mitrea et al., 2013). As the comparison of different benchmarking methods is beyond the scope of this thesis, the comparative analysis of the three HIV studies can serve as an alternative proxy for evaluating the accuracy of new methods. A more sensitive hit selection method applied to all three studies would lead to greater overlap of hits between the studies while an approach with higher specificity would lead to more statistical significance in the overlap of the comparative studies.

1.7 Bioinformatic solutions for robust hit selection from high-throughput studies

1.7.1 Gene-set and pathway centric analysis of high-throughput hits

Suites of bioinformatic tools have been suggested as auxiliary analysis methods to buttress the prioritization of hits from high-throughput studies beyond scores and ranking (Birmingham et al., 2009; Tseng et al., 2012). These advancements have shifted the practices for hits prioritization in primary high-throughput studies from the original one-step scoring and cutoff approach (Boutros & Ahringer, 2008) to more complex analysis "pipelines" that integrate enrichment and network bioinformatics to prioritize or deprioritize hits for follow-up secondary screening (B. Dutta et al., 2016; Huang et al., 2008).

A widely used bioinformatic approach is the application of enrichment analysis that selects hits not based on original scores but rather based on the enrichment of predetermined "gene sets". Enrichment analysis, also called pathway analysis, has been proposed as a way to correct for false positives (Creixell, 2015), as randomly selected false-positive hits are less likely to come from the same pathway. Various grouping methods have been proposed as a way to apply the enrichment analysis approach, such as using preset Gene Sets (Subramanian et al., 2005), Eigengenes based on coexpression and correlation patterns (Langfelder & Horvath, 2007), Metagenes based on the clustering and dimensionality reduction of coexpression data (Brunet et al., 2004), and pathway structures based on curated gene lists of known shared function (Kanehisa & Goto, 2000). Curated pathway databases such as the Kyoto Encyclopedia for Genes and Genomes (KEGG) provide updated and easy to interpret pathway gene groups (Kanehisa et al., 2017) and are in wide use for high-throughput data analysis.

The analysis method by which to apply the knowledge from these databases has gone through a critical evolution over the past ten years. The first generation of pathway analysis

methods relied on calculating the representation of pathway hits in a given list by use of a statistical calculation such as the Fisher's Exact Test or χ^2 test (L. D. Fisher, 1993; Ronald Aylmer Fisher, 1960). These approaches to interpretation and prioritization of high throughput screens by pathway analysis have been generally summarized as Over Representation Analysis (ORA), wherein the number of hits that are also found in the predetermined pathway gene set are compared to what you would expect to find according to the null hypothesis (Beißbarth & Speed, 2004; Goeman & Bühlmann, 2007). Enrichment and prioritization by ORA also have a set of limitations that an emerging class of second-generation pathway analysis approaches seek to address. Chief among these limitations is the way that the ORA enrichment evaluates the presence or absence of a pathway gene as a binary of either being in the set or being outside the set. Statistical calculation by ORA also treats all genes as being independent from each other and does not consider how their expression correlate inside versus outside the set. Functional Class Sorting (FCS) methods emphasize coordinated changes in the group of gene from the predetermined set (i.e. pathway or functional group). Statistical approaches such as the Kolmogorov-Smirnov statistic and the Wilcoxon rank sum function factor in the correlation among the genes in the pathway and look at the scores of all the pathway genes, not just the "hits" (Barry et al., 2005; Subramanian et al., 2005). ORA and FCS approaches, however, rely on the assumption that all genes in a pathway are equal indicators of enrichment. A third generation of enrichment analysis approaches were developed that consider the topology of a pathway and not merely the list of genes that make up its components. Utilizing curated and annotated databases such as Reactome (Joshi-Tope et al., 2003) or KEGG (Kanehisa et al., 2017) topology-based (TB) approaches such as ScorePAGE (Rahnenführer et al., 2004), PathNet (Bhaskar Dutta et al., 2012), and CePA (Gu et al., 2012) consider the relationships, functions, overlap, and -crucially- the centrality of different proteins in each pathway in assigning weights to the enrichment.

The reliance on pathway databases as a way to prioritize hits from high-throughput studies, however, comes with significant analytical tradeoffs irrespective of the chosen analysis method. For example, pathway databases do not cover the complete genome. An analysis I did of the KEGG database found that only 36% of human protein coding genes are curated in at least one pathway of biological processes (Figure 1.9). Exclusive reliance on pathway databases for high-throughput hit selection therefore limits the number of novel genes that can be identified, obviating one of the most important rationales for performing unbiased genome-scale screens.

Despite the plethora of alternative solutions, ORA remains widely used in the routine reporting of high-throughput studies (Dong et al., 2016). The appeal of this enrichment approach, in addition to filtering out low likelihood hits, lies partially in its intuitive interpretability; seeing the mechanisms and biological processes that a new dataset points towards is an easier launching ground from which to generate hypotheses about mechanisms, rather than trying to interpret a long list of often obscure gene names. This feat, however, is achieved by an abstraction of the data that removes the output from the units it was designed to measure and ultimately aims to validate. Current best practices in pathway analysis relegate these forms of high-throughput screen interpretation as "exploratory add-ons" (Sedeno-Cortes & Pavlidis, 2014) and highlight the challenge of converting insights on the combined gene set level back to individual gene analyses that can be followed up experimentally (Mooney & Wilmot, 2015). A challenge remains to utilize the gene set enrichment analysis in a way that still provides a path to the characterization of specific genes and mechanisms novel to the context being investigated.

1.7.2 Network database approaches to hit selection of high-throughput studies

Complementary to the pathway analysis filtering approach, the utilization of protein-protein interaction (PPI) databases have been proposed as a way to prioritize lower scoring hits from high-throughput studies and expand the dataset (B. Dutta et al., 2016; Tu et al., 2009; L. Wang et al., 2009). Based on combined curated and unbiased omic scale studies, PPI databases have broader coverage (>80%) of the genome than curated databases (Figure 1.10). In the network analysis approach, PPI databases such as the Search Tool for the Retrieval of Interacting Genes (STRING) (Szklarczyk et al., 2010), are used to find lower scoring candidates that have predicted interactions with higher scoring hits. The hit selection set is then expanded to include the lower scoring interacting hits.

Integrating the network information from interaction databases into a prioritization pipeline for high-throughput studies can be done by different approaches. The simplest to apply is the Direct Neighbor approach whereby proteins that have a direct predicted interactions with candidates of interest are considered part of the network (Oti et al., 2006). As the Direct Neighbor approach casts a very wide net alternative and more discriminating approaches to expanding candidate lists by use of network database knowledge have been created. The network propagation method uses a diffusion approach which has the benefit of penalizing nodes (in this case proteins) that have very high rates of predicted interactions (Cowen et al., 2017). Other approaches incorporate concepts from graph theory and information theory such as considering the mutual dependency between two nodes and collapsing layers of information into kernel matrices to be collectively considered (Yu et al., 2013). Where more multi-level datasets are available, network prioritization using more sophisticated statistical and machine learning methods such as linear regression and random forest have yielded more discriminating results (L. Wang et al., 2018; W. Zhang et al., 2017).

Similar to the various criteria for pathway analysis discussed in the previous section, the first generation of network analysis methods still remains in wide use, as shown by the popularity of databases like STRING which employs a direct neighbor method. Expansion of the lists of candidates from a high-throughput functional study by any of the above network approaches is often in the first steps of candidate prioritization. A network analysis approach to hit selection decreases the rate of false negatives as it expands the hit selection set to more hits. Network analysis, however, also amplifies the noise in the hit selection set as false-positive candidates in the original high scoring set of hits also expand to include their predicted interactions.

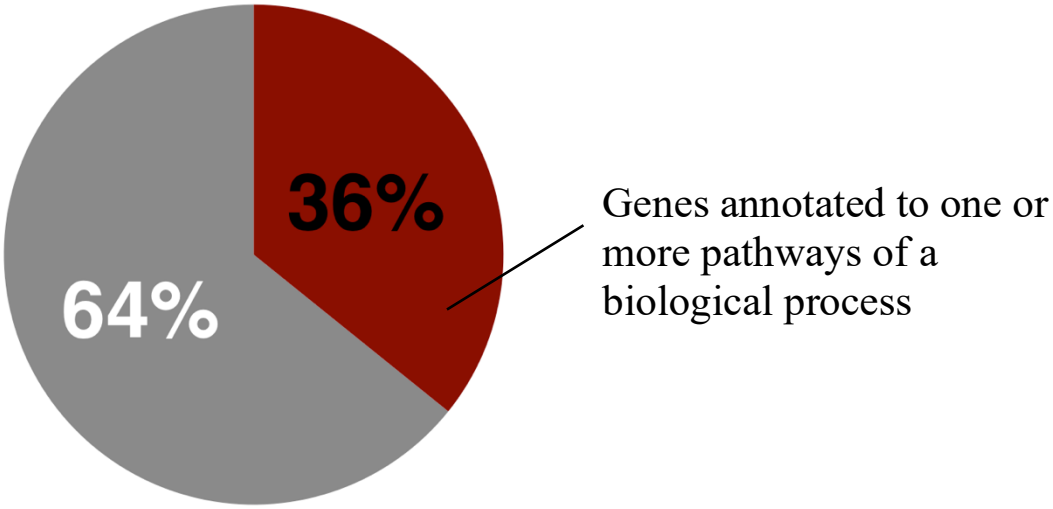


Figure 1.9: **Percentage of protein coding human genes annotated into pathways by KEGG.**

KEGG database collected on May 11, 2019, filtered for pathways relating to biological processes.
Human protein coding genes as defined by HGNC as of June 29, 2018

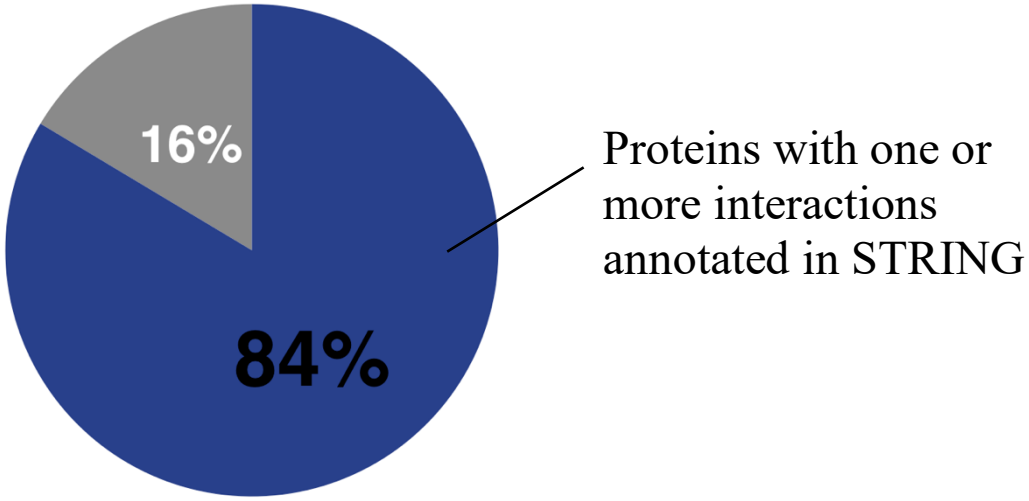


Figure 1.10: **Percentage of protein coding human genes with at least one associated interaction in the STRING database.**

STRING database collected on October 3, 2018, filtered for interactions of 400 confidence score or higher
Human protein coding genes as defined by HGNC as of June 29, 2018

1.7.3 Gene-level vs. pathway-level insights in omics output and visualization

Comparing the different analytical advancements in high throughput hit prioritization by unbiased scoring, pathway analysis, and network analysis highlights their complementary contributions to error correction, interpretability, and novel candidate identification. While frequently applied in parallel, there is currently no outlined framework for how to integrate their analytical contributions. This presents a critical challenge in applying robust hit selection to studies such as RNAi high throughput assays of the macrophage response to LPS.

1.7.4 Developing web-based versus Rscript solutions for high-throughput hit selection

A solution to the challenges of hit prioritization as outlined above would have implications for hit prioritization in other high-throughput studies, and would be of use to the screening community outside of the field of innate immunity. Critical to the utility of a bioinformatic analysis pipelines is its adaptation to platforms that broaden its access. Access and dissemination of bioinformatic solutions largely fall under two categories, making the code available through public repositories where users can download the code or, alternatively, adapting the code into a user-friendly web interface that is publicly accessible. These two approaches both have advantages and tradeoffs. Using a software repository platform has the advantage of ensuring the user that their data is safe as the code they are running only runs locally on the machines of their choosing. It also further expands the possibility for advanced developers to adapt the analysis pipeline to the specific analysis of their project. A consequence of this path is that it limits the access to only those with a basic-to-strong proficiency of working with the computational language and platform, further creating a gulf between the informaticians and the experimentalists running the assays. Building an interactive web application circumvents the accessibility challenge and substantially broadens the community

of potential users. This benefit, however, comes with intrinsic tradeoffs, most prominently the risk of data security for the users who have to upload their data to a server they do not control. Of lesser concern, but still critical to consider, web applications are less adaptable and offer more confined solutions to pre-set problems, thus limiting the applicability of the pipeline. These considerations are critical to consider when designing novel bioinformatic frameworks.

1.8 Aims

In this thesis my aim is to develop and present better ways to utilize, prioritize, and interpret data from high-throughput studies and to apply these insights to genome-scale studies of the response to LPS signaling in macrophages.

Building on the knowledge outlined in this introduction, my work herein focuses on five aims:

***Aim 1*:** To understand how regulatory candidates selected by ancillary bioinformatic and secondary validation approaches improved on hits selected based on simple ranking from high-throughput studies (Chapter 3). I use the three siRNA studies of the response to LPS and the three studies of HIV Dependency Factors (HDFs) to compare overlap and enrichment of hits selected based on bioinformatic validation versus score ranking methods.

***Aim 2*:** To test the improvements and trade-offs of pathway and network based hit prioritization methods for high throughput studies, and to develop an optimized approach that brings together their combinatorial benefits (Chapter 4). I use the three studies of HDFs as my testing data sets, the KEGG and STRING databases for my respective pathway and network databases, and hypergeometric and random permutation tests for statistical evaluation.

***Aim 3*:** To develop the new analysis approach from aim 2 into a publicly available interface so that it can be utilized for analysis of different genome-scale datasets (Chapter 5). I use the Shiny R platform to design the interface and add features for combined network and pathway visualization and data exploration.

***Aim 4*:** To apply the developed framework from aim 3 to the analysis of the three siRNA studies of the response to LPS, and to identify novel regulatory mechanisms of macrophage activation (Chapter 6). I use the analysis methods developed in the previous chapters and follow up the findings with studies using chemical inhibitors targeting the cellular processes identified by the analysis.

***Aim 5*:** To broadly characterize post-transcriptional dynamics and alternative splicing in the macrophage response to LPS by RNA-seq (Chapter 7). I use short and long read RNA-seq approaches of four macrophage cell types treated with LPS.

2 Materials and Methods

2.1 Datasets

2.1.1 Three genome-wide siRNA studies of the response to LPS

The genome-scale siRNA studies of the macrophage response to LPS used in this thesis were published in (J. Sun et al., 2017) and (Ning Li et al., 2017).

2.1.2 Genome-wide CRISPR study of the response to LPS in mouse

The CRISPR based genome-wide study of the LPS response in bone marrow derived dendritic cells used in this thesis was published by (Parnas et al., 2015). The dataset with complete Z scores and FDR was generously shared by Aviv Regev.

2.1.3 Three genome-wide siRNA studies of HIV Dependency Factors

The three genome-wide siRNA studies of essential proteins in early HIV infection were published by (Brass et al., 2008); König et al. (2008); (Zhou et al., 2008). The complete datasets of scores and metadata of these studies were generously shared by Amy Espeseth (Zhou *et al* screen), Abraham Brass (Brass *et al* screen), and Sumit Chanda (König *et al* screen).

2.1.4 RNA-seq of WT and UBL5 KD macrophages treated with LPS

Wild type THP1 cells and UBL5 knockdown cells (expressing a UBL5 specific shRNA) were differentiated into a macrophage-like state with 5ng/ml phorbol-12-myristate-13-acetate (PMA, from Sigma, P1585) for 72 hours. Following differentiation cell were treated with 100ng/mL LPS for 0h, 0.5h, 1h, 2h, and 4h. Total RNA for each sample was prepared using the TRIzol reagent (Invitrogen Cat# 15596026) and using the manufacturers protocol for

isolating RNA. Two biological replicates were generated for each time point and each condition.

The RNA-seq was run on an Illumina HiSeq2000 at a sequencing depth of 2x150bp reads, at around 20-30 million reads per sample. The results were mapped using STAR aligner as described by Dobin et al. (2013). The data was then analyzed for differential exon usage by DEXseq as described by Anders et al. (2012).

2.2 Databases

2.2.1 KEGG database for pathway enrichment

The KEGG database was downloaded from the KEGG Application Program Interface (API), as described previously (Kanehisa et al., 2017). For the analysis described in this thesis, the KEGG data was downloaded on May 11, 2019.

Pathway lists were filtered for pathways that are related to biological processes (and excluding the ones related to disease) by only selecting pathways with PathwaysIDs of 05000 or less. EntrezIDs were added to the NCBI gene symbols in the KEGG database by the org.Hs.eg.db: R package (Marc Carlson (2018). org.Hs.eg.db: Genome wide annotation for Human. R package version 3.7.0.) and the org.Mm.eg.db: R package (Marc Carlson (2018). org.Mm.eg.db: Genome wide annotation for Mouse. R package version 3.7.0.). The annotated pathway enrichment document was formatted into a matrix of gene IDs and pathway identifiers (Figure 2.1) to be subset into 2x2 matrices for competitive enrichment analysis as previously described (Goeman & Bühlmann, 2007).

EntrezID	GeneSymbol	PathwayID	PathwayName
10	NAT2	232	Caffeine metabolism
10	NAT2	983	Drug metabolism - other enzymes
10	NAT2	1100	Metabolic pathways
100	ADA	230	Purine metabolism
100	ADA	1100	Metabolic pathways
1000	CDH2	4514	Cell adhesion molecules (CAMs)
10000	AKT3	1521	EGFR tyrosine kinase inhibitor resistance
10000	AKT3	1522	Endocrine resistance
10000	AKT3	1524	Platinum drug resistance
10000	AKT3	4010	MAPK signaling pathway
10000	AKT3	4012	ErbB signaling pathway
10000	AKT3	4014	Ras signaling pathway
10000	AKT3	4015	Rap1 signaling pathway
10000	AKT3	4022	cGMP-PKG signaling pathway
10000	AKT3	4024	cAMP signaling pathway
10000	AKT3	4062	Chemokine signaling pathway
10000	AKT3	4066	HIF-1 signaling pathway
10000	AKT3	4068	FoxO signaling pathway
10000	AKT3	4071	Sphingolipid signaling pathway
10000	AKT3	4072	Phospholipase D signaling pathway
10000	AKT3	4140	Autophagy - animal

Figure 2.1: **Format of pathway enrichment file from KEGG.**

Pathway membership data was downloaded from the KEGG database and formatted into a matrix of gene IDs (EntrezID, GeneSymbol) matched with their membership group name (PathwayID, PathwayName).

2.2.2 *STRING database for biological network interactions*

The STRING database was downloaded from the STRING API as described by (Szklarczyk et al., 2010). The 9606.protein.links.full.v10.5 was downloaded for human interactions and the 10090.protein.links.full.v10.5 for mouse interactions. Inferred interactions from other species were not included. The network downloads were separated based on the evidence source of their interactions. The evidence source categories followed the STRING database categorizations; neighborhood, co-occurance, fusion, co-expression, experiments, databases, and text mining. The different evidence source network files were then split into three groups based on their evidence scores, 0.15-0.4 as low confidence, 0.4-0.7 as medium confidence, and 0.7-1 as high confidence. The files were then converted into the igraph format using the igraph R package (Csardi G, Nepusz T: The igraph software package for complex network research, InterJournal, Complex Systems 1695. 2006. http://igraph.org). For the analysis described in this thesis, the STRING data was downloaded on October 3rd, 2018. Each analysis was performed using a single master igraph that was generated by combining the igraphs of the relevant criteria (evidence source and scores). The networks were used to prioritize lower scoring hits by using the direct neighbor functional approach as previously described (L. Wang et al., 2009).

2.2.3 *Gene and protein ID conversion*

Gene to protein ID conversions were done using the biomaRt R package (Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).) EntrezID to GeneSymbol ID conversions were done using the org.Hs.eg.db: R package (Marc Carlson (2018). org.Hs.eg.db: Genome wide annotation for Human. R package

version 3.7.0.) and the org.Mm.eg.db: R package (Marc Carlson (2018). org.Mm.eg.db: Genome wide annotation for Mouse. R package version 3.7.0.)

2.2.4 Human and mouse protein coding gene lists

Lists of human protein coding genes were accessed from the HUGO Gene Nomenclature Committee (HGNC) website, using the HGNC_genes_with_protein_product_EntrezID_geneSymbol list. Lists of protein coding genes in mouse were downloaded from the Mouse Genome Informatics (MGI) database, using the MGI_MRK_Coord.rpt download and filtering for Marker Type = protein coding gene.

2.2.5 Mouse and human gene orthologues

Gene ID orthologous were curated from the biomarRt R package (Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. Steffen Durinck, Paul T. Spellman, Ewan Birney and Wolfgang Huber, Nature Protocols 4, 1184-1191 (2009).) using the mouse_human_orthologs_biomart download, last accessed on January 23, 2019. Hits within the three LPS screens that did not have an orthologue listed were checked in the NCBI Gene database, if an orthologue was listed it was added to the list. If a gene was listed as having a one-to-many orthologue mapping from human to mouse the GeneCards database (Stelzer et al., 2016) was consulted and the orthologue with the highest similarity score was used as a match for the gene.

2.2.6 Canonical TLR pathway genes

A list of canonical TLR pathway genes was curated by combining the canonical gene list from (J. Sun et al., 2016) and (Mertins et al., 2017). The combined list includes 186 genes in human and mouse IDs.

2.3 Statistics

2.3.1 Hypergeometric test for pathway enrichment

Hypergeometric distributions to calculate the significance of shared enrichments (across screens and for pathway analysis) were done by generating contingency matrices of shared and non-shared hits and then analyzing by a one sided Fishers' Test with the alternative hypothesis set to "greater than" and the null being no shared enrichment. This approach has been previously described as the "competitive enrichment" test or over representation analysis (Khatri et al., 2012).

2.3.2 Normalization of high-throughput readouts

Normalization of scores from high-throughput studies was performed using the Z score approach described in (Birmingham et al., 2009). Where plate information was available, scores were normalized to plate mean, otherwise scores were normalized to overall mean of the data set.

2.3.3 Cell viability correction

Cell viability correction for candidates from high-throughput studies was different for different studies and based on available data as follows:

THP1 -TNF -α screen: the ubiquitin promoter driven renilla luciferase readout was used as a normalization factor which can correct for cases where siRNAs target genes that affect general transcription, and also cases where specific gene knockdown affected the cell number and viability of the sample. The TNF promoter driven firefly activity provided a measure of the

TLR4 ligand induced TNF activation. The ratio of firefly to renilla luminescence (R1/R2) was used as a cell viability/general transcription corrected readout of TNF promoter activity. (Reporter cell line described in section 2.5)

Raw G9 – NF-κB and Raw G9 – TNF-α screen: The NF-κB and TNF-α screens in RAW G9 cells were provided with cell count number for each target. In the screening assay two imaging fields were collected from each well providing imaging data for approximately 300-400 cells per well. Gene candidates that had less then 50 cells per image were removed. (Reporter cell line described in section 2.5)

Zhou et al HDF Screen: The study calculated a Percent Cell Viability measure for each gene target. The values were normalized using a normal distribution and gene candidates with a cell viability score of -2 or less were flagged.

Brass et al HDF Screen: The study included cell count number for each gene target. The counts were log10 normalized and then given a plate-by-plate Z-score normalization. Gene candidates with a cell viability score of -2 or less were flagged.

König et al HDF Screen & Parnas et al CRISPR-LPS Screen: Data shared with us for these studies already had a cell viability correction applied to their readout scores.

2.4 Bioinformatics

2.4.1 Iterative analysis of pathway and network enrichment (TRIAGE)

Iterative analysis in excel

To integrate and iterate pathway and network analysis when prioritizing screen hits I began by utilizing the Comprehensive Analysis for RNAi Data (CARD) platform (B. Dutta et al., 2016). CARD is a publicly available website that curates and combines established algorithms and methods for high-throughput data pre-processing and filtering. CARD takes uploads from RNAi screens and provides a platform for data normalization and setting a cutoff. Following data normalization, the user can select from a series of further analysis steps including pathway analysis and network analysis. CARD uses a direct neighbor approach in network analysis. The user provides two cutoffs, one for "hits" and one for "non-hits" that can be prioritized based or predicted interactions from selected databases. For pathway analysis, CARD uses a Fisher's Exact Test for enrichment of pathways from different databases (KEGG, Reactome, Gene Ontology). CARD generates a series of graphs and tables based on the different analysis that the user selected. These analyses can be downloaded as separate excel files.

A critical first question to address in the design of an integrated and iterative analysis approach, is whether the iterative analysis ultimately converges on a single set of selected hits such that the cycles of iteration can terminate. As an initial attempt to test the convergence of the iterative approach, and for an early model of TRIAGE analysis, I relied on the CARD analysis features. I used data from the *THP1-TNF-α screen* (see section 2.1.1). Each step of the iterative analysis was run as a separate new project in CARD, with the results downloaded and reformatted as the input for the next step in the analysis. The step by step processes were as follows.

Preliminary Step: Prior to uploading the screen to CARD the candidates from the screen were divided into three groups and prepared as an input file for CARD.

1. Hits from the screen that were above the RNAiCut recommended threshold (Kaplow et al., 2009) (or, alternatively a very conservative threshold), that had also passed analysis by CARD for low likelihood of being an Off-target effect (Marine et al., 2012) and had not been

ruled out by expression analysis as being absent in the cell line used, were assigned as “top hits” by placing a value of 1 in the “Replicate1” column of the input file.

2. Hits that were above the chosen threshold but had been flagged by either Off-target analysis or Expression Analysis, and hits that were below the threshold but had a Z-score of -1 or lower were assigned as “medium confidence hits” by placing a value of 0.5 in the “Replicate1” column of the input file.

3. All other candidates in the screen were assigned as “non-hits” by placing a value of 0 in the “Replicate1” column of the input file

Step 1: The input file was uploaded to CARD using the “No normalization” setting of the “load data” tab. Pathway Analysis (using KEGG) was then run by setting a cutoff of 0.9. (A cutoff of 0.9 was chosen to ensure that all the hits assigned a score of 1 are selected as CARD uses a “greater than” approach to its input.) The results were locally downloaded.

Step 2: Using the output file “KEGG.enrichment.csv” the pathways with a p-value of 0.05 or less were selected. The generated list of pathways was used to filter a comprehensive list of all the pathway names from KEGG with their associated member gene names generating a list of genes associated with the selected pathways. The newly generated list of gene names from significantly enriched pathways was cross-referenced with the original input file from the screen. In a new column titled “Replicate1” gene names that had a score of 0.5 or higher in the input file that were also in the pathway generated list of genes were assigned a score of “1”. Genes that had a score of 0.5 or higher that were not in the pathway list were assigned a score of 0.5. Genes that had a score of 0 in the input file were given the same score of 0 throughout the analysis. This is the “contractive step”, the hit list is contracted to consist only of hits that make up the strongly represented pathways and processes in the set.

Step 3: The document with the newly assigned scores was uploaded to a new project in CARD using the “no normalization” option in the “load data” tab. Network analysis was performed using the combined network databases provided by CARD (HPRD, BioGRID, BIND) with the cutoff for hits set at 0.9 and the cutoff for non-hits set at 0.4. The analysis was locally downloaded.

Step 4: Using the “NormalizedDataAll.csv” file from the CARD analysis download in step 3, gene names that had a “NetworkDegree” value of greater than 0 and an input score in step 2 of 0.5 or greater were assigned a value of 1 in a new “Replicate1” column. All other gene names were assigned the same value as they were given in step 2. This is the Expansion step, where the set is expanded to include any hit with a predicted network interaction with any gene in the “contractive set” from step 2. This also completed the end of the first iteration.

Step 5: The new document is uploaded to CARD and put through steps 2-4 as the second iteration.

Step 6: After the iteration is complete, the gene names that are assigned a value of 1 at the end of the current iteration are compared with the gene names that were assigned a value of 1 at the end of the previous iteration. If the lists are not the same, the iteration is repeated by going through steps 5-6 again. If the lists are the same (i.e. the analysis has “converged” on a single set and additional iterations will keep yielding the same results) the analysis terminates.

Step 7: Once the analysis terminates the hits that are assigned as having a value of 1 in the final iteration are chosen as the newly selected hits from the screen.

Applying the above approach to the *THP1-TNF-α* study led to the analysis going through 4 iterations, after which the data converged on a set of hits with each subsequent iteration. These analyses provided initial validation of the method and supported the analysis principle that, through iterative analysis of pathway and network enrichment, a dataset of high-confidence and medium confidence hits converges on a single set of hits. The cumbersome processes, however, generally took upward of two days for each screen. To further develop and robustly test the TRIAGE analysis design, it was critical to develop a more streamlined and automated computational approach to iterative analysis.

Iterative analysis in R

Computational analysis in this thesis was done in the R environment (R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/. Genomic analysis software was supported by the BioConductor platform (Gentleman et al., 2004). To build the TRIAGE analysis pipeline in R, all components were built as separate functions and then integrated together into a master function. Below is a summary of the individual steps taken followed by their integration into an iterative function.

Importing and standardizing databases: The downloads from pathway databases and network databases (KEGG and STRING) were mapped to common IDs (EntrezIDs, see section 2.2.3 of this chapter for ID conversion). The network database was converted to a set of igraphs (section 2.2.2) and the pathway database was converted to a two-column table of pathway name and pathway members. This enabled efficient mapping between hit datasets and databases.

Pathway enrichment function: A pathway enrichment function was created that creates a contingency matrix for each pathway name in the pathway database and a list of IDs separated into "hits" and "non-hits". Using a one sided Fisher's exact test, the *p*-values, FDR, and Bonferroni correction of enrichment for each pathway name is generated. This analysis loops over all the unique pathway names in the pathway database table. The pathway function is provided with a significance cutoff (<0.055). The function uses the significance to separate the pathway names that passed the threshold of significance for the list of pathway names and creates a list of selected pathways. Using the list of selected pathways, a vector of unique gene IDs that are members of the selected pathways is created. When using a dataset with a single cutoff (hits vs. non-hits), the intersect of the pathway member IDs and hits IDs is sub selected as a new set of hits. When using a dataset with two cutoffs (high confidence hits, medium confidence hits, and non-hits), only the high confidence hits are used as the "hits" for determining pathway enrichment. After the vector of pathway associated genes is created, a vector of the union of high confidence and medium confidence hits is generated. The intersect of the new vector of hits and the vector of pathway genes is taken as the new set of hits.

Network enrichment function: A generated igraph of the selected network database parameters is matched with a list of high confidence hits and medium confidence hits. A new igraph is created based on the intersect of the list of hits with the database igraph. A two-column table of each of the two IDs ("nodes") from each predicted interaction ("edge") is created. The list is then matched with a list of the high confidence hits. To find the medium confidence hits that have predicted interactions with high confidence hits, edges that have a match with the high confidence list of hits in at least one of their nodes are kept while nodes without a match in either of its edges are filtered out. A new vector of unique IDs is generated from the filtered

table and the union of the vector and the high confidence list of hits are assigned as the new set of high confidence hits.

Iterative analysis function: The iterative function was built by first creating a pipeline where the pathway enrichment function is applied to the input of a screen containing gene IDs in three groups (high confidence, medium confidence, and non-hits). The output of the pathway analysis steps is then reshaped to match the required input for the network analysis step. Following the network analysis output, the new hit characterizations are assigned as the new input column for pathway analysis. A "confidence category" column keeps track of what confidence level each hit was at the first input while a "proxy score" column updates the level of hit confidence each gene is assigned within each iteration. A separate data frame is created where the selected enrichment pathways from the pathway function are tabulated.

To halt the iterative loop when the analysis starts generating the same set of high-confidence hits across iterations, the script uses a *while* function relying on a variable nested in an *if* function. Briefly, the variable "counter" is assigned as TRUE at the start of the analysis. An additional variable ("iteration") counts what iteration of the analysis is currently running. The analysis function is wrapped within a "while" function that only runs the analysis while counter = TRUE. Following an iteration of pathway and network analysis, an *if* function evaluates if the iteration count is greater than 1. If true, the *if* function evaluates if the table of IDs with associated proxy score of this iteration match the table of IDs with associated proxy scores from the previous iteration. If the condition is true the "counter" variable is assigned as counter = FALSE. This leads to the termination of the function. Otherwise, the counter variable remains TRUE and the condition of the *while* loop is met to commence a new iteration of the analysis. When the analysis is complete a data frame with all the input IDs, the confidence category each ID was assigned at input, the proxy score for each iteration, and whether the ID

was assigned as a "hit" by the final iteration is generated. An additional data frame with the pathway enrichments from the final iteration, each pathway name matched with the intersect of member IDs with the list of hits IDs is also generated.

Repeated tests with different datasets have all resulted in the analysis converging to a single set of hits after a finite number of iterations. When testing randomized datasets, however, a number of datasets (out of more than five thousand tested) led to the set oscillating between different sets after a few iterations. To ensure the termination of the iterative analysis even in these rare cases, an additional condition was added to the above described test. The results of each iteration after iteration ≥ 3 are compared to the results of all the previous iterations. If a result is repeated it is indicative of an oscillating pattern. The analysis then finds the iteration within the repeated pattern that has the largest hit set and then terminates the analysis and assigns that iteration as the final output. (R code excerpted in Inset 2.1)

Adaptable TRIAGE function:

To make TRIAGE a more adaptable framework for iterative analysis with different datasets and databases, a R script version of TRIAGE was written using the inputs as variables. The TRIAGE function relies on calling two separate analysis functions, the pathway enrichment function (Inset 2.2) and the network analysis function (Inset 2.3). The master TRIAGE function applies the pathway and network functions iteratively, and the results are tested for when the analysis converges on a single set (Inset 2.4).

R code for termination of iterative analysis

```
1.  ##-- Measure if there is an iterating pattern
2.      converge.sequence <- 0
3.
4.      if (iteration >=3) {
5.        highconf.length <- c(length(append.hits.df[[ID.column]][append.hits.df[[Network
    .iteration.name]]== "HighConf"]))
6.
7.        for ( t in 1:iteration){
8.          if (identical(append.hits.df[[Network.iteration.name]], append.hits.df[[paste
    0("NETWORK.iteration_", iteration-t)]])){
9.            converge.sequence <- t
10.           break
11.         } else {
12.           highconf.length <- c(highconf.length, length(append.hits.df[[ID.column]][ap
    pend.hits.df[[paste0("NETWORK.iteration_", iteration-t)]] == "HighConf"]))
13.         }
14.       }
15.     }
16.
17.     ##-- See if Iteration is converging on output
18.
19.     if((iteration != 1 && identical(append.hits.df[[Network.iteration.name]], append.
    hits.df[[paste0("NETWORK.iteration_", iteration-1)]]))
20.         || (converge.sequence > 0
21.             && (length(append.hits.df[[ID.column]][append.hits.df[[Network.iteration.n
    ame]]== "HighConf"]) == max(highconf.length))
22.         ))  {
23.       ##-- Set counter to false
24.       counter <- FALSE
25.     } else {
26.       ##-- Update itertion number
27.       iteration <- iteration + 1
28.     }
```

Inset 2.1: **R code to ensure termination of iterative analysis.**

Set of selected hits are compared to the set of selected hits at all previous iterations (lines 7-8). If the set of selected hits is identical to the set in the previous iteration (line 19) or if the set of hits oscillates between a defined set of hits and the current iteration gives the largest set (lines 20-21) the counter is set to FALSE (line 24) leading the analysis to terminate and select the current iteration as the final output.

R code for pathway enrichment

```
1.  ########## ENRICHMENT 2 Tier FUNCTION ----
2.
3.  ###** Requirments **###
4.  ## screen.datafame: A dataframe of the screen
5.  ## ID.column: A column within the screen.dataframe for the identifiers of the targets (EntrezID, GeneSymbol, etc.)
6.  ## criteria.column: A column within the screen.dataframe of the criteria for being considered a hit
7.  ## highconf.criteria: A criteria each target has to meet to be considered a "high confidence" hit.
8.  ## midconf.criteria: A criteria each target has to meet to be considered a "mid confidence" hit.
9.  ## criteria.setting: Whether you should be using "equal", "greater than or equal", or "less then or equal". Should be i
    n the format of "equal", "greater", or "less"
10. ## enrichment.dataframe: A dataframe to be used for pathway membership in the format of a column of IDs (should be same
     as ID column in screen.dataframe in ID type and column title) and a column of which group they are part of of. (each I
    D~group relationship should be in its own seperate row)
11. ## enrichment.title: Name of the column with the names of the enrichment groups the targets are members of.
12. ## stat.test: name of of the statistical test to be used for measuring enrichment confidence. Should be in format of ei
    ther "pVal", "FDR", or "Bonferroni"
13. ## test.cutoff: A numeric value which a less than the value in stat.test will be considered a significant enrichment.
14.
15.
16. ###** Output **###
17. ## a list with 2 dataframes
18. ## 1: the screen.dataframe with an appended column listing each ID if it is part of a significantly enriched enrichment
     ("Yes") and was either high or medium confidence in the input, if it is not ("No"), or if it is missing from the enric
    hment.dataframe ("Missing")
19. ## 2: a dataframe listing all the enrichmen groups with number of members, number of hits, ID of hits, p value, FDR, an
    d Boneferroni of enrichment.
20.
21. ENRICHMENT.2tiers.function <- function(screen.dataframe, ID.column, criteria.column, highconf.criteria, midconf.criteri
    a, criteria.setting, enrichment.dataframe, enrichment.title, stat.test, test.cutoff){
22.
23.   ##--Assign Dataframes
24.   #Get dataframe of hits and assign temp column names
25.   hits.df <- screen.dataframe[, c(which(colnames(screen.dataframe) == ID.column), which(colnames(screen.dataframe) == c
    riteria.column))]
26.   names(hits.df) <- c("ID.temp", "criteria.temp")
27.
28.   #Get dataframe of enrichment and assign temp names
29.   enrich.df <- enrichment.dataframe[, c(which(colnames(enrichment.dataframe) == ID.column), which(colnames(enrichment.d
    ataframe) == enrichment.title))]
30.   names(enrich.df) <- c("ID.temp", "enrich.temp")
31.
32.   ##--Get high confidence, medium confidence hits and non hits matrix
33.   if (criteria.setting == "equal") {
34.     highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == highconf.criteria)])
35.     midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == midconf.criteria)])
36.     all.hits <- union(highconf.hits, midconf.hits) # a seperate matrix of high and low confidence hits to be used for g
    raph subsetting later
37.     non.hits <- setdiff(as.matrix(hits.df$ID.temp), all.hits)
38.   } else if (criteria.setting == "greater") {
39.     highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp >= highconf.criteria)])
40.     midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp >= midconf.criteria && hits.df$criteria.temp
    < highconf.criteria)])
41.     all.hits <- union(highconf.hits, midconf.hits) # a seperate matrix of high and low confidence hits to be used for g
    raph subsetting later
42.     non.hits <- setdiff(as.matrix(hits.df$ID.temp), all.hits)
43.   } else if (criteria.setting == "less") {
44.     highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp <= highconf.criteria)])
45.     midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp <= midconf.criteria && hits.df$criteria.temp
    > highconf.criteria)])
46.     all.hits <- union(highconf.hits, midconf.hits) # a seperate matrix of high and low confidence hits to be used for g
    raph subsetting later
47.     non.hits <- setdiff(as.matrix(hits.df$ID.temp), all.hits)
48.   } else {
49.     highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == highconf.criteria)])
50.     midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == midconf.criteria)])
51.     all.hits <- union(highconf.hits, midconf.hits) # a seperate matrix of high and low confidence hits to be used for g
    raph subsetting later
52.     non.hits <- setdiff(as.matrix(hits.df$ID.temp), all.hits)
53.     message("criteria.setting not properly defined. Using 'equals to' default")
54.   }
55.
56.   ##-- subset hits and non.hits list to only those that intersect with what is in the enrichment dataframe
57.   subset.hits <- intersect(highconf.hits, enrich.df$ID.temp)
58.   subset.all.hits <- intersect(all.hits, enrich.df$ID.temp)
59.   subset.non.hits <- intersect(union(non.hits, midconf.hits), enrich.df$ID.temp)
60.
61.   ##- subset enrichment list for enrichments that have members in the screen.dataframe list and get a list of unique na
    mes
62.   enrich.filter <- enrich.df[which(enrich.df$ID.temp %in% hits.df$ID.temp), ]
63.   enrich.unique <- unique(as.character(enrich.filter$enrich.temp))
64.
65.
66.   ##- add create columns and rows for all the pathways
67.   p.val <- enrich.group.members.number <- enrich.hit.number <- enrich.hit.IDs <- rep(NA,length(enrich.unique))
68.
69.
70.   ##-- populate the columns with the values for each enrichment group
71.
(Cont.)
```

R code for pathway enrichment (continued)

```
72.
73.    #- Fischer's Test (greater then)
74.    for(i in 1:length(enrich.unique))
75.    {
76.      enirchment.members <- enrich.filter$ID.temp[which(as.character(enrich.filter$enrich.temp) == enrich.unique[i])]
77.      contingency <- matrix(NA,nrow = 2, ncol = 2)
78.      contingency[1,1] <- length(intersect(enirchment.members, subset.hits)) # pathway.genes.hits
79.      contingency[1,2] <- length(intersect(enirchment.members, subset.non.hits)) # pathway.genes.non.hits
80.      contingency[2,1] <- length(setdiff(subset.hits, enirchment.members)) # non.pathway.hits
81.      contingency[2,2] <- length(setdiff(subset.non.hits, enirchment.members)) # non.pathway.non.hits
82.      p.val[i] <- fisher.test(contingency, alternative = "greater")$p.value
83.      enrich.group.members.number[i] <- contingency[1,1] + contingency[1,2]
84.      enrich.hit.number[i] <- contingency[1,1]
85.      enrich.hit.IDs[i] <- paste(unique(enrich.filter$ID.temp[match(intersect(enirchment.members,subset.hits), enrich.fil
    ter$ID.temp)]),collapse = ", ")
86.    }
87.
88.    p.val.FDR <- p.adjust(p.val,method = "BH") #Correction for multiple testing
89.    p.val.FWER <- p.adjust(p.val,method = "bonferroni") #Bonferroni Correction
90.
91.
92.    #- print number of enrichment groups being calculated
93.    message(paste0("unique enrichment groups being measured: ", length(enrich.unique)))
94.
95.    ##-- Create enrichment results dataframe
96.
97.    enrichment.results <- data.frame(Enrichment = enrich.unique,
98.                                     pVal = p.val,
99.                                     pValFDR = p.val.FDR,
100.                                    pValBonferroni = p.val.FWER,
101.                                    EnrichmentMembers = enrich.group.members.number,
102.                                    EnrichmentHitNumber = enrich.hit.number,
103.                                    EnrichmentHitID = enrich.hit.IDs)
104.
105.   enrichment.results <- enrichment.results[with(enrichment.results, order(enrichment.results$pValBonferroni,enrichment.
    results$pValFDR,enrichment.results$pVal)),] #order dataframe from lowest to highest value of statistical test
106.
107.   ###--
    - Assign individual IDs in the screen.dataframe if they are hits based on the enrichment analysis and the provided cuto
    ff
108.
109.   ##-- Get list of enrichment that are siginificantly enriched for
110.   if (stat.test == "pVal") {
111.     sig.enrichment <- as.character(enrichment.results$Enrichment[which(enrichment.results$pVal < test.cutoff)])
112.   } else if (stat.test == "FDR") {
113.     sig.enrichment <- as.character(enrichment.results$Enrichment[which(enrichment.results$pValFDR < test.cutoff)])
114.   } else if (stat.test == "Bonferroni") {
115.     sig.enrichment <- as.character(enrichment.results$Enrichment[which(enrichment.results$pValBonferroni < test.cutoff)
    ])
116.   } else {
117.     sig.enrichment <- as.character(enrichment.results$Enrichment[which(enrichment.results$pVal < test.cutoff)])
118.     message("Statisitical test not properly defined. Unsing pVal as default")
119.   }
120.
121.   #- list of hit IDs in significantly enriched enrichments
122.   sig.enrich.hitIDs <- intersect(unique(enrich.filter$ID.temp[which(enrich.filter$enrich.temp %in% sig.enrichment)]), s
    ubset.all.hits)
123.
124.   #- list of IDs that are not members of the significantly
125.   if(length(sig.enrich.hitIDs > 0)){
126.     nonhits.sig.enrich.IDs <- setdiff(enrich.filter$ID.temp, sig.enrich.hitIDs)
127.   } else {
128.     nonhits.sig.enrich.IDs <- enrich.filter$ID.temp
129.   }
130.
131.   ##-
    - append column to screen.dataframe of whether the row ID is annotated in the enrichment.dataframe and if it is or isn'
    t part of a significantly enriched group
132.   temp.enrich.IDs <- matrix("Missing", nrow(hits.df))
133.   temp.enrich.IDs[hits.df$ID.temp %in% sig.enrich.hitIDs] <- "Yes"
134.   temp.enrich.IDs[hits.df$ID.temp %in% nonhits.sig.enrich.IDs] <- "No"
135.
136.   screen.dataframe$Enrichment.hit <- temp.enrich.IDs
137.
138.
139.
140.   ###--- Define return objects
141.   enrichment.output <- list(screen.dataframe, enrichment.results)
142.
143.   ####---- Return
144.   return(enrichment.output)
145.
146. }
```

Inset 2.2: **R code for stand-alone pathway enrichment function.**

The R script enables the selection of hits from a dataset with high confidence and medium confidence hits based on competitive pathway enrichment. The R script can be used with any user provided pathway membership resource and the user can select the preferred statistical test and cutoff.

R code for enrichment by network analysis

```
1.  ########## NETWORK FUNCTION ----
2.
3.  ###** Requirments **###
4.  ## screen.datafame: A dataframe of the screen
5.  ## ID.column: A column within the screen.dataframe for the identifiers of the targets (EntrezID, GeneSymbol, etc.).

6.  ## criteria.column: A column within the screen.dataframe of the criteria for being considered a high confidence hit,
     medium confidence hit, and low confidence/non hit.
7.  ## highconf.criteria: A criteria each target has to meet to be considered a "high confidence" hit.
8.  ## midconf.criteria: A criteria each target has to meet to be considered a "mid confidence" hit.
9.  ## criteria.setting: Whether you should be using "equal", "greater than or equal", or "less then or equal". Should b
    e in the format of "equal", "greater", or "less".
10. ## network.igraph: an igraph of the network to be used for network analysis (network igraph must use the same ID typ
    e as screen.dataframe)
11.
12.
13. ###** Output **###
14. ## The input screen.dataframe with two appended columns
15. ## Network.analysis: A column with information on whether it was an "InputHighConfidenceHit" or "NetworkAnalysisAdde
    d", if neither than "0" is assigned.
16. ## Network.hit: A column on wether based on network enrichment the row ID is a hit, with designations "Yes" and "No"

17.
18. NETWORK.function <- function(screen.dataframe, ID.column, criteria.column, highconf.criteria, midconf.criteria, crit
    eria.setting, network.igraph){
19.
20.   ##-- necessary libraries
21.   library("igraph")
22.
23.   ##--Assign Dataframes
24.   #Get dataframe of hits and assign temp column names
25.   hits.df <- screen.dataframe[, c(which(colnames(screen.dataframe) == ID.column), which(colnames(screen.dataframe) =
    = criteria.column))]
26.   names(hits.df) <- c("ID.temp", "criteria.temp")
27.
28.   ##--Get high confidence, medium confidence hits and non hits matrix
29.   if (criteria.setting == "equal") {
30.     highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == highconf.criteria)])
31.     midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == midconf.criteria)])
32.     all.hits <- union(highconf.hits, midconf.hits) # a seperate matrix of high and low confidence hits to be used fo
    r graph subsetting later
33.     non.hits <- setdiff(as.matrix(hits.df$ID.temp), all.hits)
34.   } else if (criteria.setting == "greater") {
35.     highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp >= highconf.criteria)])
36.     midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp >= midconf.criteria && hits.df$criteria.te
    mp < highconf.criteria)])
37.     all.hits <- union(highconf.hits, midconf.hits) # a seperate matrix of high and low confidence hits to be used fo
    r graph subsetting later
38.     non.hits <- setdiff(as.matrix(hits.df$ID.temp), all.hits)
39.   } else if (criteria.setting == "less") {
40.     highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp <= highconf.criteria)])
41.     midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp <= midconf.criteria && hits.df$criteria.te
    mp > highconf.criteria)])
42.     all.hits <- union(highconf.hits, midconf.hits) # a seperate matrix of high and low confidence hits to be used fo
    r graph subsetting later
43.     non.hits <- setdiff(as.matrix(hits.df$ID.temp), all.hits)
44.   } else {
45.     highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == highconf.criteria)])
46.     midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == midconf.criteria)])
47.     all.hits <- union(highconf.hits, midconf.hits) # a seperate matrix of high and low confidence hits to be used fo
    r graph subsetting later
48.     non.hits <- setdiff(as.matrix(hits.df$ID.temp), all.hits)
49.     message("criteria.setting not properly defined. Using 'equals to' default")
50.   }
51.
52.   ###--- Format igraph
53.   set.seed(123)
54.
55.   ##-- Select subgraph that matches list of IDs in screen  ----> This will be the geneal all encompasing network
56.   G <- upgrade_graph(network.igraph)
57.   OverallDegree <- degree(G)
58.   screen.IDs.for.network.analysis <- intersect(hits.df$ID.temp, V(G)$name)
59.   Graph <- induced.subgraph(G, screen.IDs.for.network.analysis)
60.
61.   ##-- Select subgraph that matches list of IDs in high confidence AND mid confidence hits ----
    > This will be the network of all "hits"
62.   Subset.allhitIDs.for.network.analysis <- intersect(all.hits, V(G)$name)
63.   allhits.SubGraph <- induced.subgraph(G, Subset.allhitIDs.for.network.analysis)
64.   allhits.SubGraph <- induced.subgraph(allhits.SubGraph, names(which(igraph::degree(allhits.SubGraph) > 0)))
65.   allhits.SubGraph.IDs <- V(allhits.SubGraph)$name
66.
67.   ##-- Removes edges that don't connect a high confidence hit.
68.
69.   graph.edges.allhits.names <- get.data.frame(allhits.SubGraph, what = "edges")
70.   indices.to.remove <- intersect(which((graph.edges.allhits.names$from %in% highconf.hits) == FALSE),
71.                                  which((graph.edges.allhits.names$to %in% highconf.hits) == FALSE))
72.
73.   if(length(indices.to.remove) > 0){
74.     graph.edges.allhits.names <- graph.edges.allhits.names[-indices.to.remove, ]
75.   }
76.
77.
(Cont..)
```

R code for enrichment by network analysis (continued)

```
  ##-- Get net list of added High confidence hits
    new.highconf.SubGraph <- graph.data.frame(graph.edges.allhits.names, directed = FALSE, vertices = NULL)

    ##--Combine this list with the original high confidence hits
    new.highconf.IDs <- union(V(new.highconf.SubGraph)$name, highconf.hits)

    ###-- Create columns to append to screen.dataframe
    hits.df$Network.analysis <- 0
    hits.df$Network.analysis[match(new.highconf.IDs, hits.df$ID.temp)] <- "NetworkAnalysisAdded"
    hits.df$Network.analysis[match(highconf.hits, hits.df$ID.temp)] <- "InputHighConfidenceHit"
    hits.df$Network.hit <- "No"
    hits.df$Network.hit[match(new.highconf.IDs, hits.df$ID.temp)] <- "Yes"

    ##-- append to the screen.dataframe input
    screen.dataframe.out <- screen.dataframe
    screen.dataframe.out$Network.analysis <- hits.df$Network.analysis
    screen.dataframe.out$Network.hit <- hits.df$Network.hit



    ###--- Return the appended data frame inside a list
    network.output <- list(screen.dataframe.out)
    return(network.output)

}
```

Inset 2.3: **Rcode for stand-alone network analysis function.**

The R script enables the selection of hits from a dataset with high confidence and medium confidence hits based on direct neighbor interactions. The R script can be used with any user provided interaction network.

R code for TRIAGE analysis

```
1.  ########## TRIAGE FUNCTION ----
2.
3.
4.  ###** Libraries **###
5.  library('dplyr')
6.  library('data.table')
7.  library('igraph')
8.
9.
10. ###** Requirments **###
11. ## screen.datafame: A dataframe of the screen
12. ## ID.column: A column within the screen.dataframe for the identifiers of the targets (EntrezID, GeneSymbol, etc.)
13. ## criteria.column: A column within the screen.dataframe of the criteria for being considered a hit
14. ## highconf.criteria: A criteria each target has to meet to be considered a "high confidence" hit.
15. ## midconf.criteria: A criteria each target has to meet to be considered a "mid confidence" hit.
16. ## criteria.setting: Whether you should be using "equal", "greater than or equal", or "less than or equal". Should be i
    n the format of "equal", "greater", or "less"
17. ## enrichment.dataframe: A dataframe to be used for pathway membership in the format of a column of IDs (should be same
     as ID column in screen.dataframe in ID type and column title) and a column of which group they are part of of. (each I
    D~group relationship should be in its own seperate row)
18. ## enrichment.title: Name of the column with the names of the enrichment groups the targets are members of.
19. ## stat.test: name of of the statistical test to be used for measuring enrichment confidence. Should be in format of ei
    ther "pVal", "FDR", or "Bonferroni"
20. ## test.cutoff: A numeric value which a less than the value in stat.test will be considered a significant enrichment.
21. ## network.igraph: an igraph of the network to be used for network analysis (network igraph must use the same ID type a
    s screen.dataframe)
22.
23.
24. ###** Output **###
25. ## Output is a list of 3 dataframes
26. ## [[1]] input dataframe plus 'TRIAGE.hit' column,
27. ## [[2]] dataframe of high confidence and medium confidence designation at each iteration,
28. ## [[3]] dataframe of final TRIAGE enrichments
29.
30.
31. TRIAGE.2tiers.function <- function(screen.dataframe, ID.column, criteria.column, highconf.criteria, midconf.criteria, c
    riteria.setting, enrichment.dataframe, enrichment.title, stat.test, test.cutoff, network.igraph) {
32.
33.   ###--- Save original inputs
34.   input.screen.dataframe <- screen.dataframe
35.   input.ID.column <- ID.column
36.   input.criteria.column <- criteria.column
37.   input.highconf.criteria <- highconf.criteria
38.   input.midconf.criteria <- midconf.criteria
39.   input.criteria.setting <- criteria.setting
40.   input.enrichment.dataframe <- enrichment.dataframe
41.   input.enrichment.title <- enrichment.title
42.   input.stat.test <- stat.test
43.   input.test.cutoff <- test.cutoff
44.   input.network.igraph <- network.igraph
45.
46.
47.
48.   ###--- Define input high confidence hits, medium confidence hits, and background
49.   #Get dataframe of hits and assign temp column names
50.   hits.df <- screen.dataframe[, c(which(colnames(screen.dataframe) == ID.column), which(colnames(screen.dataframe) == c
    riteria.column))]
51.   names(hits.df) <- c("ID.temp", "criteria.temp")
52.
53.   ##-- Get high confidence, medium confidence hits and non hits matrix of the inputs
54.   if (criteria.setting == "equal") {
55.     input.highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == highconf.criteria)])
56.     input.midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == midconf.criteria)])
57.     input.all.hits <- union(input.highconf.hits, input.midconf.hits) # a seperate matrix of high and low confidence hit
    s to be used for graph subsetting later
58.     input.non.hits <- setdiff(as.matrix(hits.df$ID.temp), input.all.hits)
59.   } else if (criteria.setting == "greater") {
60.     input.highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp >= highconf.criteria)])
61.     input.midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp >= midconf.criteria && hits.df$criteria
    .temp < highconf.criteria)])
62.     input.all.hits <- union(input.highconf.hits, input.midconf.hits) # a seperate matrix of high and low confidence hit
    s to be used for graph subsetting later
63.     input.non.hits <- setdiff(as.matrix(hits.df$ID.temp), input.all.hits)
64.   } else if (criteria.setting == "less") {
65.     input.highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp <= highconf.criteria)])
66.     input.midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp <= midconf.criteria && hits.df$criteria
    .temp > highconf.criteria)])
67.     input.all.hits <- union(input.highconf.hits, input.midconf.hits) # a seperate matrix of high and low confidence hit
    s to be used for graph subsetting later
68.     input.non.hits <- setdiff(as.matrix(hits.df$ID.temp), input.all.hits)
69.   } else {
70.     input.highconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == highconf.criteria)])
71.     input.midconf.hits <- as.matrix(hits.df$ID.temp[which(hits.df$criteria.temp == midconf.criteria)])
72.     input.all.hits <- union(highconf.hits, input.midconf.hits) # a seperate matrix of high and low confidence hits to b
    e used for graph subsetting later
73.     input.non.hits <- setdiff(as.matrix(hits.df$ID.temp), input.all.hits)
74.     input.message("criteria.setting not properly defined. Using 'equals to' default")
75.   }
76.
(Cont.)
```

R code for TRIAGE analysis (continued)

```
77.     ###--- Begin setting the first iteration variable to 1
78.     iteration <- 1
79.
80.     #- Create dataframe for appending the dataframe through each iteration step
81.     append.hits.df <- hits.df
82.     names(append.hits.df)[which(colnames(append.hits.df) == "ID.temp")] <- input.ID.column  ## Necesary to rename the ID
      column to original name so that it aligns with the ID.column of the enrichment dataframe
83.
84.     ####---- Set up a counter for while loop till iteration converge
85.     counter <- TRUE
86.
87.     while (counter == TRUE) {
88.
89.       ###--- Enrichment step ----> Contracting the hits set
90.
91.       enrichment.step <- ENRICHMENT.2tiers.function(screen.dataframe, ID.column, criteria.column, highconf.criteria, midc
      onf.criteria, criteria.setting, enrichment.dataframe, enrichment.title, stat.test, test.cutoff)
92.
93.       ##-- create enrichment output dataframe
94.       #- Get datframe from enrichemnt step
95.       enrichment.output.df <- enrichment.step[[1]]
96.
97.       ##-- Append Data frame to be used for Network function input
98.       append.hits.df <- data.frame(append.hits.df, temp = "", stringsAsFactors = FALSE)
99.       Enrichment.iteration.name <- paste0("ENRICH.iteration_", iteration)
100.      names(append.hits.df)[names(append.hits.df) == "temp"] <- Enrichment.iteration.name
101.      append.hits.df[[Enrichment.iteration.name]][enrichment.output.df$Enrichment.hit == "Yes" & (append.hits.df[[ID.colu
      mn]] %in% input.all.hits)] <- "HighConf"
102.      append.hits.df[[Enrichment.iteration.name]][enrichment.output.df$Enrichment.hit != "Yes" & (append.hits.df[[ID.colu
      mn]] %in% input.all.hits)] <- "MedConf"
103.
104.
105.      ###--- Network step ----> Expanding the hits set
106.
107.      ##-- Set Network Function input parameters
108.      screen.dataframe <- append.hits.df
109.      ID.column <- ID.column
110.      criteria.column <- Enrichment.iteration.name
111.      highconf.criteria <- "HighConf"
112.      midconf.criteria <- "MedConf"
113.      criteria.setting <- "equal"
114.      network.igraph <- network.igraph
115.
116.
117.      ##-- Run NETWORK function
118.
119.      network.output <- NETWORK.function(screen.dataframe, ID.column, criteria.column, highconf.criteria, midconf.criteri
      a, criteria.setting, network.igraph)
120.      network.output.df <- network.output[[1]]
121.
122.      ##-- Append Data frame to be used as Enrichment input or final output
123.      append.hits.df <- data.frame(append.hits.df, temp = "", stringsAsFactors = FALSE)
124.      Network.iteration.name <- paste0("NETWORK.iteration_", iteration)
125.      names(append.hits.df)[names(append.hits.df) == "temp"] <- Network.iteration.name
126.      append.hits.df[[Network.iteration.name]][network.output.df$Network.hit == "Yes" & (append.hits.df[[ID.column]] %in%
       input.all.hits)] <- "HighConf"
127.      append.hits.df[[Network.iteration.name]][network.output.df$Network.hit != "Yes" & (append.hits.df[[ID.column]] %in%
       input.all.hits)] <- "MedConf"
128.
129.      ##-- Set Enrichment Function input parameters for next iteration
130.      screen.dataframe <- append.hits.df
131.      ID.column <- input.ID.column
132.      criteria.column <- Network.iteration.name
133.      highconf.criteria <- "HighConf"
134.      midconf.criteria <- "MedConf"
135.      criteria.setting <- "equal"
136.      enrichment.dataframe <- enrichment.dataframe
137.      enrichment.title <-  enrichment.title
138.      stat.test <- stat.test
139.      test.cutoff <- test.cutoff
140.
141.      ##-- Print message on completion of iteration
142.      message(paste("iteration ", iteration, " Complete"))
143.
144.      ##-- Measure if there is an iterating pattern
145.      converge.sequence <- 0
146.
147.      if (iteration >=3) {
148.        highconf.length <- c(length(append.hits.df[[ID.column]][append.hits.df[[Network.iteration.name]]== "HighConf"]))
149.
150.        for ( t in 1:iteration){
151.          if (identical(append.hits.df[[Network.iteration.name]], append.hits.df[[paste0("NETWORK.iteration_", iteration-
      t)]])){
152.            converge.sequence <- t
153.            break
154.          } else {
155.            highconf.length <- c(highconf.length, length(append.hits.df[[ID.column]][append.hits.df[[paste0("NETWORK.iter
      ation_", iteration-t)]] == "HighConf"]))
156.          }
157.        }
158.      }
159.
(Cont.)
```

R code for TRIAGE analysis (continued)

```
##-- See if Iteration is converging on output

    if((iteration != 1 && identical(append.hits.df[[Network.iteration.name]], append.hits.df[[paste0("NETWORK.iteration_", iteration-1)]]))
       || (converge.sequence > 0
           && (length(append.hits.df[[ID.column]][append.hits.df[[Network.iteration.name]]== "HighConf"]) == max(highconf.length))
       ))  {
      ##-- Set counter to false
      counter <- FALSE
    } else {
      ##-- Update itertion number
      iteration <- iteration + 1
    }

  }
  ####---- end of while loop, Print message on completion of TRIAGE iteration
  message(paste("TRIAGE iterations complete, number of iterations: ", iteration))



  ####---- Create TRIAGE output dataframes
  input.plus.triage.df <- input.screen.dataframe
  input.plus.triage.df$TRIAGE.hit <- network.output.df$Network.hit

  hits.by.iteration.df <- append.hits.df
  names(hits.by.iteration.df)[which(colnames(hits.by.iteration.df) == "criteria.temp")] <- input.criteria.column

  triage.enrichment.df <- enrichment.step[[2]]

  ##-- Create List of outputs
  triage.output <- list(input.plus.triage.df, hits.by.iteration.df, triage.enrichment.df)

  #- Print message on list content
  message("list contents: \n [[1]] input dataframe with 'TRIAGE.hit' column, \n [[2]] dataframe of high confidence and medium confidence designation at each iteration, \n [[3]] dataframe of final TRIAGE enrichments")


  ####---- return list
  return(triage.output)
}

```

Inset 2.4: **R code for TRIAGE function.**

The R script enables the selection of hits from a dataset with high confidence and medium confidence hits using the iterative TRIAGE analysis. The R script calls the enrichment function and network function from Inset 2.1 and 2.2, respectively. The analysis can be performed using the user's choice of pathway database and interaction network.

The list of input variables that can be selectively assigned in the adaptable TRIAGE function in R and their required formats are:

screen.datafame: A data frame of the screen.

ID.column: A column within the screen.dataframe for the identifiers of the targets (EntrezID, GeneSymbol, etc.).

criteria.column: A column within the screen.dataframe of the criteria for being considered a hit.

highconf.criteria: A criteria each target has to meet to be considered a "high confidence" hit.

midconf.criteria: A criteria each target has to meet to be considered a "mid confidence" hit.

criteria.setting: Whether the function should be using "equal", "greater than or equal", or "less than or equal" when assessing if confidence criteria is met. criteria.setting input should be in the format of "equal", "greater", or "less".

enrichment.dataframe: A data frame to be used for pathway membership in the format of a column of IDs (should be same as ID column in screen.dataframe in ID type and column title) and a column of which group they are part of. (each ID~group relationship needs to be in its own separate row).

enrichment.title: Name of the column with the names of the enrichment groups the targets are members of.

stat.test: Name of the statistical test to be used for measuring enrichment confidence. Needs to be in the format of either "pVal", "FDR", or "Bonferroni".

test.cutoff: A numeric value which a less than value in stat.test will be considered a significant enrichment.

network.igraph: an igraph of the network to be used for network analysis (network igraph must use the same ID type as screen.dataframe).

The user provided variables are then used to apply the iterative function as in the previous paragraph. The adaptable version of TRIAGE broadens the possibility for its application beyond the use of the specific databases and settings it was originally designed with. The TRIAGE function provides an output in the format of a R script list that contains three data frames: 1) the input data frame plus a 'TRIAGE.hit' column. 2) A data frame of high confidence and medium confidence designation at each iteration of the analysis. 3) A data frame of final TRIAGE enrichments from the provided enrichment data frame.

2.4.2 Web based interface of TRIAGE (Shiny)

The TRIAGE web interface was designed to run on a set of intuitive user inputs and provide the user with results of TRIAGE analysis and the ability to explore and download the results (Figure 2.2). Creation of the public facing web page based on R script was done using the Shiny application (Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2019). shiny: Web Application Framework for R. R package version 1.3.2. https://CRAN.R-project.org/package=shiny). Briefly, the different sets of outputs were separated into different "tabs" with an additional tab added for input. Inputs required from the user were separated into "selectedInputs" (organism, pathway, network, interaction confidence for network analysis), "conditionalPanel" (selecting interaction network confidence source), "fileInput" (uploading input file), "textInput" (high-conf cutoff value, mid-conf cutoff value), "checkboxInput" (add genome background), and "actionButton" (run analysis, reset analysis). The inputs are assigned to variables that are then matched to variables in the TRIAGE function.

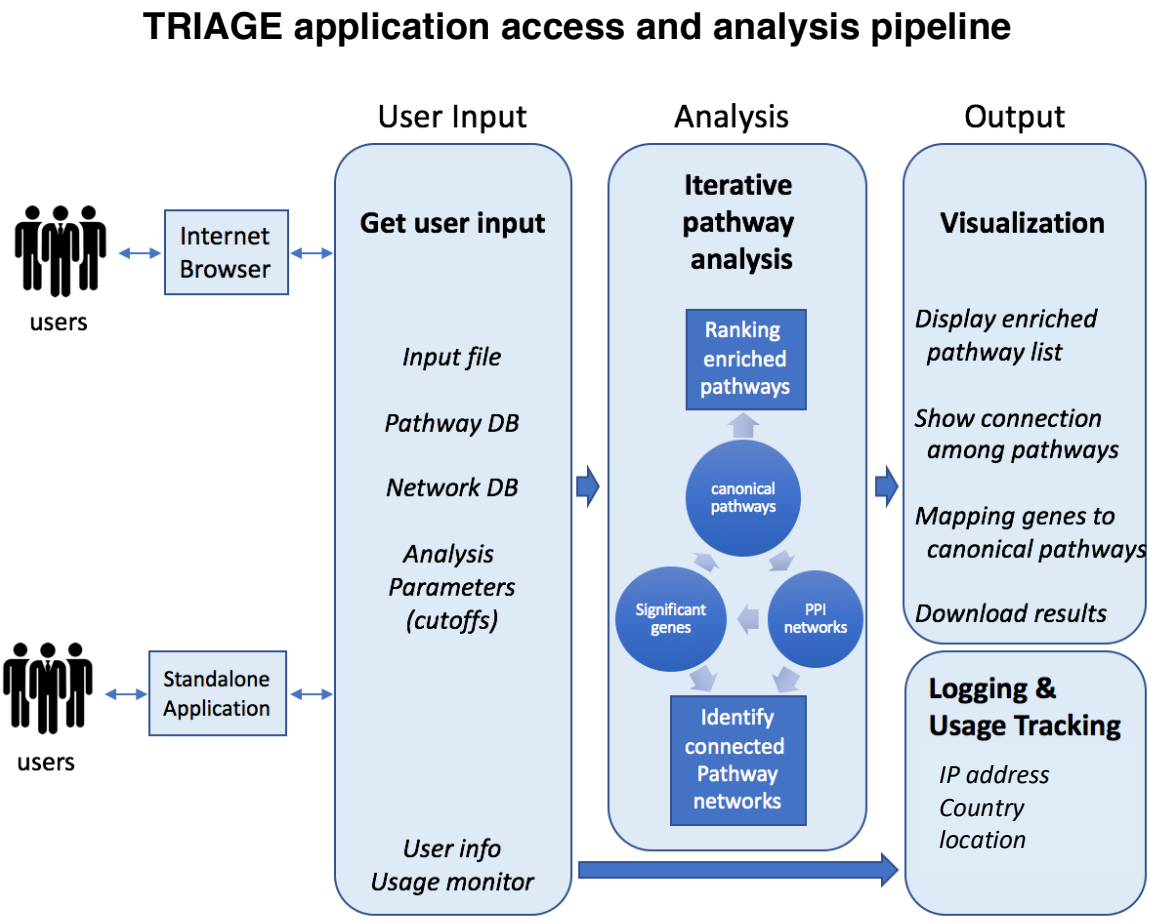


Figure 2.2: **Shiny application for TRIAGE analysis.**

Schematic of the Shiny application of the TRIAGE analysis interface. The application can be run either as a stand-alone application or accessed from an internet browser. The analysis relies on a set of user inputs (uploaded gene list with scores, cutoffs to be used, and selecting pathway and network database parameters). The application applies TRIAGE analysis to the uploaded data. The application tracks only the IP address and location of connection. No data of the analysis or uploaded screen is collected or stored by the application. The analysis output provides a download zip folder of analysis as well as an interactive interface for network and pathway exploration.

A set of warning messages were built in if lists of hits yield no results in pathway and network analysis (a warning message suggests lowering cutoff criteria and re-running the analysis). Additional warning messages were built in for organism mismatch and lack of gene ID recognition. To visualize the progress of the analysis for the user a progress bar appears and fills to 1/3 of the bar when the analysis is ready to begin and to half of the bar after the first iteration. After each iteration the progress bar fills one third of the remaining space of the bar. When the analysis is complete the progress bar fills to the end and the input tab switches to the Enriched Pathways tab for a review of the results.

To create the hyperlinks for each enriched pathway with mapped on hits on KEGG pathway maps a link2KEGGmapper function is generated following an analysis on TRIAGE. The link2KEGGmapper function generates a list of gene names mapped to the organism abbreviation and assigned colors based on input provided confidence level. A web path is created for each pathway and added to the end of the https://www.kegg.jp/kegg-bin/show_pathway?%s0%s web address. This generates a unique URL for each pathway based on the list of high confidence and medium confidence hits in its membership to match the URL generated by the KEGG mapper and ID color feature (https://www.genome.jp/kegg/tool/map_pathway3.html).

For the table of pathway enrichment an enrichment score (*EnirchScore)* for each pathway was calculated. The score is a measure of the robustness of the pathways enrichments by the number of genes represented in the TRIAGE dataset. The EnrichScore also evaluates how many of the genes driving the pathway enrichment were assigned as high confidence in the input (HighScoreGenes). The total EnrichScore is calculated as $\left(\frac{HitGenes}{GenesInPathway} + \frac{HighScoreGenes}{HitGenes}\right) / 2$.

To generate the appended columns of "InteractingGenes" and "NetworkGenePathways" for the TRIAGE gene hits tab, an igraph of the selected hits is

generated based on the network input parameters provided by the user and filtered into a sub-igraph for each hit. The interacting genes are then cross-referenced with the pathway input parameters selected by the user and the list of pathway memberships of the interacting genes are tabulated, counted, and added to the "NetworkGenePathways" column. The download tab on the interface was created as a reactive page. As files are added to the directory with additional analysis steps, the download page updates with a list of file names in the current directory. For ease of use the download files are put in a zip file format.

The application is hosted by the National Institute of Allergy and Infectious Disease (NIAID) Office of Cyber Infrastructure and Computational Biology (OCICB) at the following URL: https://triage.niaid.nih.gov. The analysis is run behind two internet security firewalls and all requests are handled using encrypted connections (Figure 2.3). Using these encrypted connections ,only the browser from the IP address where the request originated from can access the data generated and uploaded. After a connection ends the directory with the upload and analysis files are deleted from the server.

2.4.3 Interactive pathway and network exploration (JavaScript, jsons, d3.js)

Interactive visual interfaces were built by integrating the JavaScript language into the R Shiny platform. Communication across the platforms were done by creating JavaScript files in R using the jasonlite R package (Jeroen Ooms (2014). The jsonlite Package: A Practical and Consistent Mapping Between JSON Data and R Objects. arXiv:1403.2805 [stat.CO] URL https://arxiv.org/abs/1403.2805.) and then fed into d3.js file (Bostock et al., 2011).

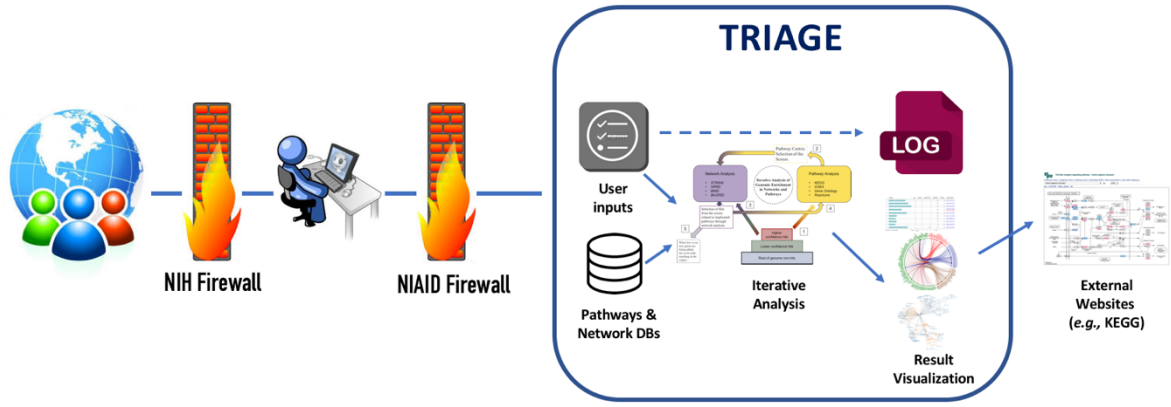


Figure 2.3: **Data security measures for accessing the TRIAGE interface.**

Access by external users to TRIAGE interface pass through two secure firewalls. The incoming request first passes through the NIH web hosting firewall after which the analysis and TRIAGE application go through the NIAID firewall where the analysis is hosted. Migrations to external websites outside of the firewall (i.e. the KEGG interface) are accompanied with a warning message for the user.

To create the hierarchical edge bundling maps of selected pathways and TRIAGE hits, an igraph of all the selected hits is generated. A vector of all the selected pathway names and additional group "novel hits" is also created. To filter the network map, first the edges are filtered based on membership in the selected pathways or interaction with an edge in one of the selected pathways. Second, nodes are filtered based on having the two edges in different groups in the vector of selected pathways and novel hits (this removes intra-group nodes). The nodes are assigned color grouping based on the edge that is in a pathway group.

The interactive interface for network exploration was built using a combination of Shiny and Java based features. Lists of clicked paths in Shiny were converted to object notation lists for Java. Communication between the json file where the clicks are tracked and the shiny interface where the clicks are received was done using d3.js. A d3.js input and output file is created for each directory to communicate between the formats (jsons for Java, csv for Shiny). Visual parameter controls of the graph on the interface are created using the Shiny slider function. A window in the interface maps the node selected by the cursor to the selected network data frame and populates the field with interaction confidence and evidence source information on the selected node. A log of clicked edges and nodes is formatted into a csv format that is added to the download directory in TRIAGE and appears in the reactive "download" tab.

2.4.4 IPA

Pathway analysis and visualization by IPA were performed using the QIAGEN Ingenuity Pathway Analysis application (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis).

2.4.5 rMATS

rMATS analysis was performed using the rMATS analysis package as described by Shen et al. (2014).

2.4.6 SQANTI

SQANTI analysis was performed using the SQANTI analysis pipeline as described by Tardaguila et al. (2018).

2.5 Cell Culture

2.5.1 Immortalized human and mouse macrophage cell-lines

RAW264.7 cells were maintained in DMEM, 10% FBS, 20mM Hepes, and 2mM glutamine. THP1 cells were maintained in RPMI1640, 10% FBS, and 2mM glutamine (containing 500ug/ml G418 for THP1 B5 line). THP1 cells were differentiated into a macrophage-like state with 5ng/ml phorbol-12-myristate-13-acetate (PMA, from Sigma, P1585) for 72 h.

2.5.2 RAW264.7 G9 cell line

The RAW264.7 G9 cell line is described in (N. Li et al., 2015). The cell line expresses enhanced green fluorescent protein (EGFP)-tagged RelA and TNF- promoter (-1229nt to -27nt)-induced mCherry.

2.5.3 THP1 – B5 cell line

The THP1 – B5 cell line is described in (N. Li et al., 2015). The cell line expresses firefly luciferase driven by the human TNF-a promoter combined with renilla luciferase driven by a ubiquitin promoter.

2.5.4 UBL5 knockdown cell line

shRNAs targeting human *UBL5* were designed using the Hannon Lab online tool (http://cancan.cshl.edu/RNAi_central/RNAi.cgi?type=shRNA) and cloned into the pEN_miRc2 vector (Shin et al., 2006). The knockdown efficiency was tested in 293T cells by transient co-expression of shRNAs with a YFP-IFIT1 fusion protein. The most efficient shRNA, targeting the following sequence in human UBL5; shRNA#1: CGGGATGAACCTGGAGCTTTAT, was subcloned to the pDS_FBneo plasmid and production of retrovirus and generation of stable cell lines were carried out as described previously (X. Zhu et al., 2007). Stable knockdown of UBL5 was confirmed using qPCR (Figure 2.4).

2.5.5 Bone marrow derived macrophages

Bone marrow progenitors isolated from sex-matched wild-type C57BL/6J mice (Jackson Laboratories) were differentiated into bone marrow derived macrophages (BMDMs) during a 6 day culture in complete Dulbecco's modified Eagle's medium (DMEM + 10% FBS, 100 U/ml penicillin, 100 U/ml streptomycin, 2 mM L-Glutamine, 20 mM HEPES) supplemented with 60 ng/ml recombinant mouse M-CSF (R and D systems). One day prior to stimulation, cells were rinsed with cold PBS, then scraped from plates using a cell lifter. 2mL of cells in complete DMEM at a concentration of $7.5x10^5$ cells/mL were then plated in 6 well plates and allowed to rest overnight at 37˚C, 5% CO2, 95% relative humidity prior to stimulation.

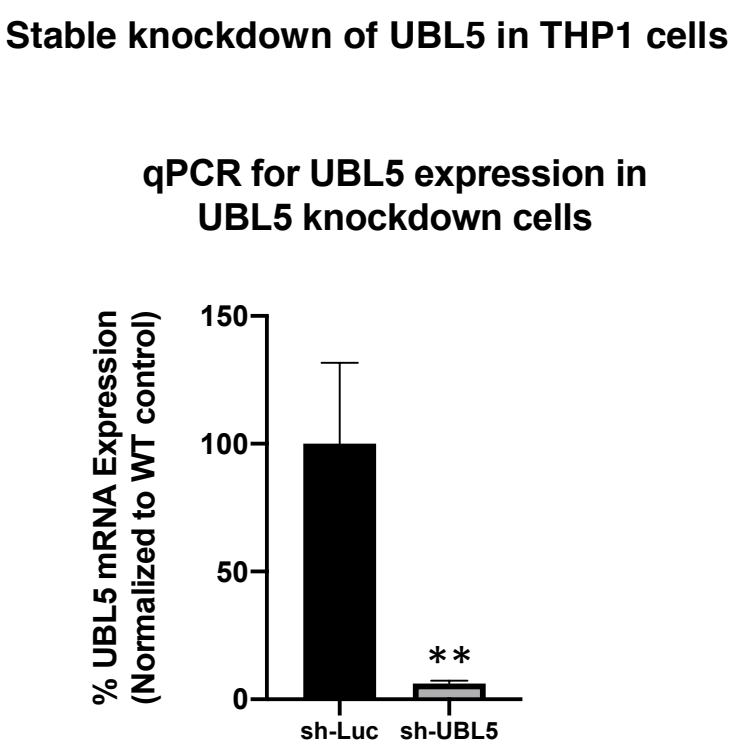


Figure 2.4 **Efficacy of UBL5 knockdown in THP1 Cells.**

Percent UBL5 mRNA expression of THP1 cells treated with a control shRNA (sh-Luciferase) and THP1 cells with UBL5 knockdown. Experiment was done in triplicates. ** = $p < 0.01$, two-tailed t test.

2.5.6 Primary human macrophages

Human blood derived monocytes were propagated in RPMI media with 10% FBS, 10 mM HEPES, and Betamercaptoethanol in 5% CO2 at 37°C. Human peripheral blood monocyte samples from screened, healthy donors were obtained under the NIH Clinical Center IRB-approved protocol 99-CC-0168 from the NIH Department of Transfusion Medicine. Human primary monocytes were differentiated with 10 ng/ml GM-CSF (R&D) for 7 days.

2.6 Assays

2.6.1 LPS treatment

LPS was from Enzo life science (used to be Alexis Biochemicals, Salmonella minnesota R595 TLRgrade, ALX-581-008-L002) and was used to treat the cells at a concentration of 100 ng/ml, unless otherwise noted.

2.6.2 Cytotoxicity

Cytotoxicity in the inhibitor experiment were determine by the release of LDH using the CytoTox 96 Non-Radioactive Cytotoxicity Assay (Promega G1780) and used according to the manufacturer's instructions. Cells were plated at $2.0x10^5$ cells per well and done in triplicates. Florescence was measured with the FLUOstar Omega Filter-based multi-mode microplate reader.

2.6.3 Splicing inhibition by Madrasin

Madrasin was from Sigma (Cat# SML 1409-5MG). To get the compound in solution, 5mG of Madrasin was diluted in 3.212mL DMSO to create a 5 mM stock. Cells were plated at $2.0x10^5$

cells per well and treated with 50ng/mL of inhibitor (except where noted differently as in the response curve assay) suspended in phenol red free media. Inhibitor was kept on for 4 hours in 37°C. Inhibitor plus media was removed and replaced with fresh media including LPS.

2.6.4 Proteasomal inhibition by MG132

MG132 was from Sigma (Cat# M7449-1ML). The cell was pre-treated with 10 μM MG132 for 4 hrs before LPS stimulation.

2.6.5 mCherry readout of TNF-α promoter activation

Raw 264.7 reporter cells were plated into a 96-well plate at $2x10^5$ cells/200μL/well with phenol red-free DMEM supplemented with 10% Hyclone® FCS, 2mM L-glutamine and 20mM HEPES, and then cultured overnight. The fluorescent intensity of mCherry expressed in the cells was measured at 584/620BP12 nm (excitation/emission filter) with the FLUOstar Omega Filter-based multi-mode microplate reader.

2.6.6 PCR

Total RNA was extracted with the RNeasy Mini Kit (QIAGEN). cDNA was reverse transcribed from 1μg RNA using the BioRad Reverse Transcription kit. PCR reactions were performed in an Applied Biosystems Veriti 96 Well Thermocycler with the following thermal cycles 95°C for 3 min, (95°C for 15 s, 55°C for 15 s, 72°C for 30 s) × 35 cycles. The samples were run on 2% agarose gels.

2.6.7 qPCR

Total RNA was extracted with the RNeasy Mini Kit (QIAGEN). cDNA was reverse transcribed from 1µg RNA using the BioRad Reverse Transcription kit. 200 ng RNA equivalent of cDNA was used per reaction with gene specific primers and FAM conjugated probes (IDT DNA) and qPCR Solaris mix (Dharmacon/Life Technologies). qPCR reactions were performed in a BioRad PCR thermocycler (BioRad CFX Real-Time Systems) with the following thermal cycles 95°C for 15 min, (95°C for 15 s, 60°C for 60 s) × 40 cycles. The Ct values were analyzed with BioRad software. Primers used in all PCR and qPCR reactions are listed in table 2.1.

Target	Oligonucleotide
Human TNF-α, Forward *(sense)*	*CCAGGGACCTCTCTCTAATCA*
Human TNF-α, Reverse *(antisense)*	*TCAGCTTGAGGGTTTGCTAC*
Human TNF-α, FAM Probe	*AGTGACAAGCCTGTAGCCCATGTT*
Mouse TNF-α, Forward *(sense)*	*CCCTCCAGAAAAGACACCATG*
Mouse TNF-α, Reverse *(antisense)*	*GTCTGGGCCATAGAACTGATG*
Mouse TNF-α, FAM Probe	*ACCGATCACCCCGAAGTTCAGTAGA*
Human HPRT, Forward *(sense)*	*CTGGAAAGAATGTCTTGATTGTGG*
Human HPRT, Reverse *(antisense)*	*CTTGCGACCTTGACCATCTT*
Human HPRT, FAM Probe	*AGACTTTGCTTTCCTTGGTCAGGCA*
Mouse HPRT, Forward *(sense)*	*TCCTCCTCAGACCGCTTT*
Mouse HPRT, Reverse *(antisense)*	*TTCCAAATCCTCGGCATAATGA*
Mouse HPRT, FAM Probe	*CCCAGCGTCGTGATTAGCGATGAT*
Human β-actin, Forward *(sense)*	*TGAAGTCTGACGTGGACATC*
Human β-actin, Reverse *(antisense)*	*ACTCGTCATACTCCTGCTTG*

IRAK1 S1, Forward *(sense)*	*TGGTCAGAGGGCTGTGAAGAC*
IRAK1 A1, Reverse *(antisense)*	*AGCCAGACCTGCTTGCAGTG*
IRAK1 S2, Forward *(sense)*	*TTGAGAAGCACCCAGAGCAC*
IRAK1 A2, Reverse *(antisense)*	*TGGAGTCAAGTGCCAGGAG*
IRAK1 S3, Forward *(sense)*	*CGCAGATTATCATCAACCC*
IRAK1 A3, Reverse *(antisense)*	*CATCAGCTCTGAAATTCATCAC*
TMED7-TICAM2 S1, Forward *(sense)*	*TGACAACGCCAAGCAGTG*
TMED7-TICAM2 A1, Reverse *(antisense)*	*ACAAAGGTGGGTCTTCTCCAAC*
TMED7-TICAM2 S2, Forward *(sense)*	*TGAGGGACTGTCCAAGAAAG*
TMED7-TICAM2 A2, Reverse *(antisense)*	*AATCGATGACAGACTTCAGAGC*
TMED7-TICAM2 A3, Reverse *(antisense)*	*CTGTGAGTCAGGGGTTAATG*
TMED7-TICAM2 S3, Forward *(sense)*	*CTGGGCCAGAAAGGAAGAC*
TMED7-TICAM2 A4, Reverse *(antisense)*	*TCCTGGACTTGTATCCACACTG*

Table 2.1 **Primer design for PCR assays.**

2.6.8 RNA extraction

Cells were plated in 10cm dish at $4.0x10^7$ cells per plate. Following LPS stimulation at the relevant time point media was removed. Cells were washed twice with PBS. 1mL of lysis buffer (100:1 Buffer RLT, β-mercaptoethanol) was added to the plate to lyse the cells. Plates were kept in -80°C till RNA extraction assay. Total RNA was extracted with the RNeasy Mini Kit (Qiagen, 74106) following the manufacturer's instruction.

2.6.9 Western Blot

THP1 cells were lysed in radioimmunoprecipitation assay buffer (Sigma, R0278) containing a protease inhibitor cocktail (Roche). Cell lysates were quantified by protein assay (Bio-Rad), and equal protein amounts were resolved with a 4 to 12% Bis-Tris Gel/MOPS Running Buffer System (Invitrogen) and transferred to nitrocellulose membranes. The membranes were analyzed by Western blotting with the following antibodies: Anti-IkBa (CST, Cat#: 4814), anti-phosphor-Erk (CST, Cat# 9106S), anti-MAPK p38, phospho (Thr180/Tyr182) (Cell Signaling, Cat# 4511).

2.6.10 Short read RNA-seq

3 replicates of each condition were prepared and tested for RNA integrity number (RIN) of 9.6 or higher. Samples were pooled and sequenced on HiSeq using Illumina TruSeq Stranded mRNA Library Prep and paired-end sequencing. Reads of the samples were trimmed for adapters and low-quality bases using Trimmomatic software before alignment with the reference genome (hg38 for human and Mouse - mm10 for mouse) and the annotated transcripts using STAR. Gene expression quantification analysis was performed for all samples using STAR/RSEM tools.

2.6.11 Long read RNA-seq

Representative samples for all conditions were selected from the short-read RNA-seq sample preparation and sequenced by PacBio Whole-Genome De Novo Sequencing. Readouts were analyzed by the SQANTI analysis pipeline.

3 Global analysis of three genome-scale siRNA studies of the LPS response in macrophages

3.1 Introduction

As I have argued in chapter 1, the sheer scale of the transcriptional response to LPS suggests that critical regulatory mechanisms of the TLR4 signal transduction pathway remain to be discovered. Robust hit selection from high-throughput gene-perturbation studies provides a means through which novel regulatory candidates can be identified. Hit selection that goes beyond a handful of familiar or highest scoring hits, however, faces significant challenges of false positives and missed hits. In this chapter I show how comparative analysis of genome-scale studies of the LPS response reflects the challenges of robust hit selection from high-throughput screens. I further demonstrate the ways in which current ancillary hit selection approaches succeed or fail to correct for the biases and error rates in high scoring hits. These analyses provide a roadmap for what novel bioinformatic solutions should seek to solve.

To go through these steps, I start with an analysis of the three genome-scale siRNA studies in macrophages by Li and Sun (Ning Li et al., 2017; J. Sun et al., 2017). I apply normalization and cell toxicity corrections to generate a comparative list of high scoring putative positive and negative regulators (section 3.2). I show how enrichment of canonical TLR pathway genes in the highest scoring hits from each study reflect the assay and readout design of each study (section 3.3). Comparing the highest scoring hits across the three studies of LPS, I show how the limited overlap suggests high rates of both false positives and false negatives in the current hit selection sets (section 3.4). The limited overlap is similarly observed when comparing the hit selection sets from the three siRNA screens and the hits selected by another genome-scale CRISPR/Cas9-based screen of the LPS response (section 3.5). To compare the overlap of reported hits from high-throughput studies and how they relate to the highest scoring hits from those studies, I also analyze three studies of HDFs for HIV (section 3.6). I show how current hit selection approaches from high-throughput studies have the

strongest impact on false positive correction when compared to hits selected purely by screen score. I also demonstrate that the described alternative approaches to hit selection still lack a correction mechanism to promote lower scoring hits that are likely potential regulators of the process being studied (section 3.7). Finally, I summarize the steps I took to methodically analyze the data from the high-throughput studies of the LPS response to identify the limitations of current approaches.

3.2 Normalization and hit selection from three siRNA studies identifies extensive lists of putative regulatory targets for macrophage activation

To begin a global analysis of regulatory candidates of the TLR4 signaling pathway, I utilized recently generated datasets in the Fraser lab from three genome-scale siRNA studies of the LPS response in macrophages. These datasets, with input from the analysis I describe here, were subsequently published in *Scientific Data* (see publications arising from this thesis). The three studies included the THP1 TNF-α study described by Sun *et al.* 2017 (J. Sun et al., 2017), and the Raw G9 TNF-α and Raw G9 NFκB studies described by Li *et al.* 2017 (Ning Li et al., 2017). (A detailed description of these three studies are in the introduction to this thesis, section 1.5.2). To select the highest scoring hits from each screen I first applied robust normalization to the readout scores. I used a robust Z score approach which normalizes the data distribution as deviations from the mean on a plate by plate basis (reviewed in section 1.5.4 and described in Dutta *et al,* 2016 (B. Dutta et al., 2016)). Ensuring that low readouts from the assay are not driven by a low number of surviving cells (potentially driven by the essentiality of gene targeted by the siRNA) required different approaches in different studies. To correct for cell viability (or general perturbation of transcription) in the THP1 study, I divided the readouts from the TNF-α promoter-driven firefly luciferase reporter by the ubiquitin promoter-driven renilla luciferase readouts, thus ensuring the samples with low cell counts have the firefly

luciferase readout corrected. In the RAW G9 cells I used a cutoff of 50 cells within the imaging field from the high content imaging assay to remove candidates that fell below it. (I describe the respective cell viability correction methods in detail in section 2.3.3). These analyses generated three datasets with normal distributions centered around a mean of 0 (Figure 3.1A).

Following score normalization and cell viability correction I applied a ranking to each gene target based on its score. To interpret the readouts from the assays as putative regulators of the TLR pathway, siRNA targets with high ranking negative scores were considered candidates with a positive regulatory role in LPS signaling, siRNA targets with high ranking positive scores were considered candidates for a putative negative regulatory role. For initial hit selection I used the top 2.5 percentile in both directions to assign as high scoring hits. This approach led to a Z-score cutoff of -2.00 with 453 hits for putative positive regulators from the THP1 TNF-α screen, a cutoff of -1.86 with 418 hits for the RAW G9 NF-κB hits, and a cutoff of –1.46 with 419 hits for the RAW G9 TNF-α screen (Figure 3.1B). In the opposite direction this approach assigned a cutoff of 1.575 with 453 hits for putative negative regulators from the THP1 TNF-α screen, a cutoff of 2.022 with 418 hits for the RAW G9 NF-κB negative regulators hits, and a cutoff of 2.397 with 418 hits for the RAW G9 TNF-α screen (Figure 3.1C). The parameters, controls, and results from normalization applied to the three screens are tabulated in Table 3.1.

Hit selection from 3 siRNA studies of the response to LPS

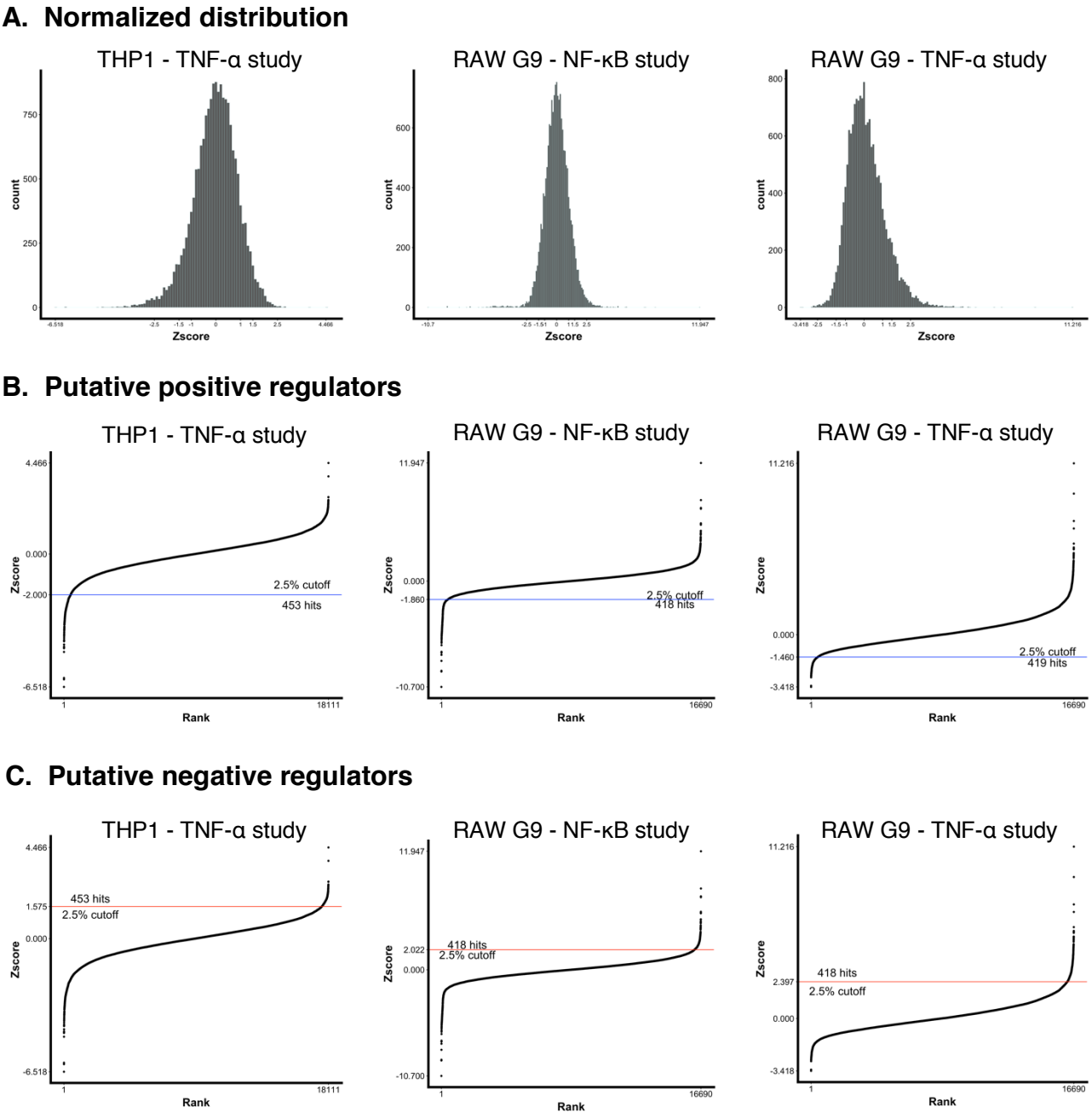


Figure 3.1: **Normalization and hit selection from 3 genome-wide studies of the macrophage response to LPS.**

(A) Normalized distribution of the readout scores for the THP1-dual luciferase study (left), RAW G9 – GFP tagged NF-κB assay (center), and RAW G9 – mCherry tagged TNF-α assay (right). (B) Selecting negative scoring targets as putative positive regulators by assigning a cutoff for Zscores in the top 2.5% of negative scores. (C) Selecting positive scoring targets as putative neagtive regulators by assigning a cutoff for Zscores in the top 2.5% of positive scores.

	THP1 TNF- α	Raw G9 TNF- α	RAW G9 NF-κB
Library	GE Dharmacon Human siGENOME SMARTpool siRNA Library Refseq27	Dharmacon siGENOME siRNA mouse library	Dharmacon siGENOME siRNA mouse library
Gene Targets	18,110	16,870	16,870
Essential-Gene Correction	Ubiquitin promoter-driven renilla luciferase activty	Cell Count > 50	Cell Count > 50
Negative Controls	2 Non-Targeting Control (NTC)	2 NTC, *Ppib*	2 NTC, *Ppib*
Positive Controls	*TLR4, IRAK1, IKBKG, MAP3K7, Renilla* siRNA	*Tlr4, Myd88, Ikbkg, Irak1,* siGFP siRNA	*Tlr4, Myd88, Ikbkg, Irak1,* siGFP siRNA
Assay Readout	*firefly/renilla* dual luciferase assay	mCherry High Content Imaging	GFP High Content Imaging
Positive Regulator Cutoff	-2	-1.86	-1.46
Positive Regulator Hits	453	418	419
Negative Regulator Cutoff	1.575	2.022	2.397
Negative Regulator Hits	453	418	418

Table 3.1 **Normalization and hit selection from three siRNA studies of the macrophage response to LPS.**

3.3 Shared and divergent enrichment for canonical members of the TLR4 pathway in hits from three LPS studies

Using the bioinformatic platform Ingenuity Pathway Analysis (QIAGEN Inc., https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis) I analyzed the selected hits from the three screens to measure their enrichment for canonical components of the TLR signaling pathway. The analysis (Figures 3.2-3.4) showed robust enrichment for critical nodes in the TLR signaling pathway in all three studies. The enrichment of expected essential effectors in the downstream LPS response represented in the high scoring selected hits set, such as MyD88 and NF-κB related genes, highlight the efficacy of the three screens. The divergence in effector enrichments between the three screens also relate to the differences

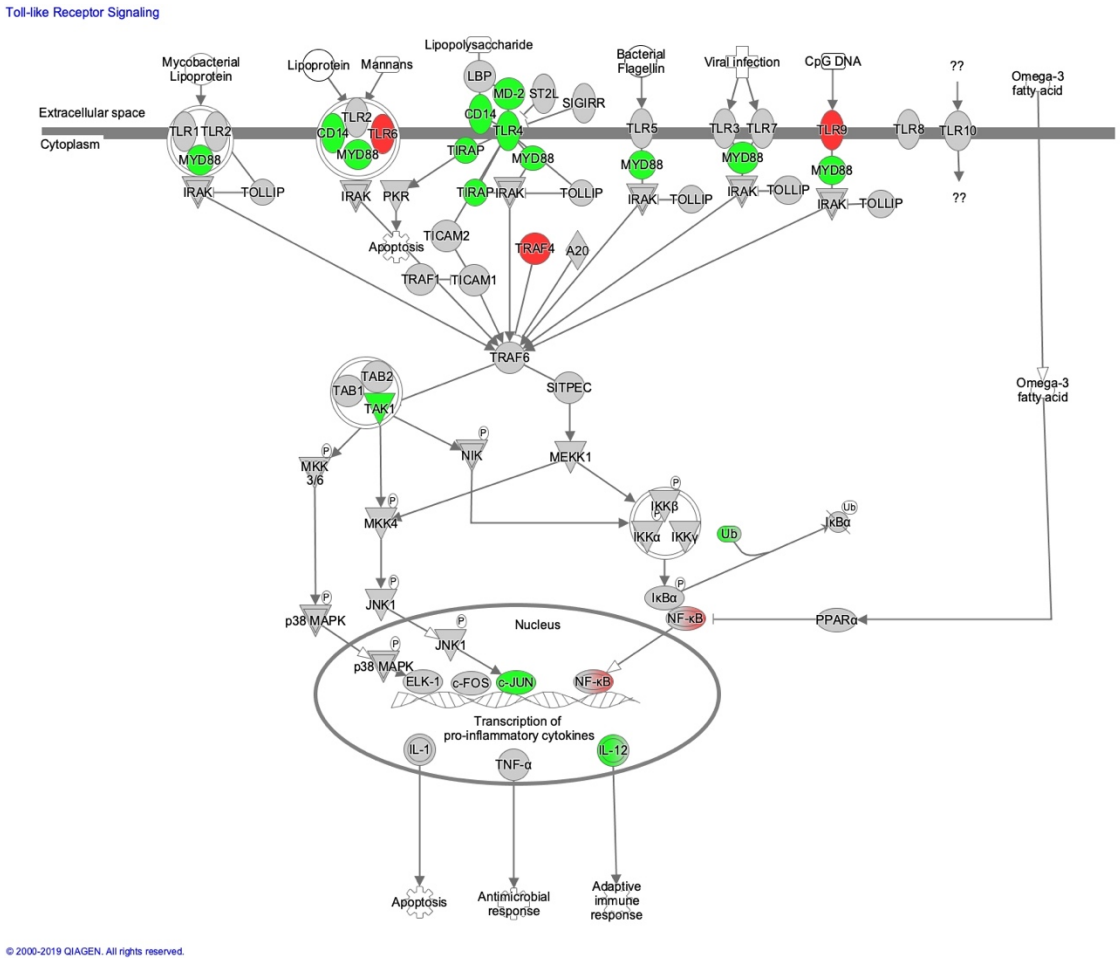
in the screen assays. For example, components of the TAK1/TAB complex, whose downstream effectors do not rely on NF-κB, appear as more critical in the two screens using the TNF-α readout, while core NFκB components are more prominent hits in the screen using the NF-κB readout (Figure 3.2-3.4). (An unexpected result was the appearance of IKKß as negative regulator in the screen using the NF-κB readout (Figure 3.3), though it may be due to an off-target effect.)

3.4 Overlap across parallel screens of the response to LPS is significant but limited

As described in chapter 1 (section 1.6.2), measuring the commonality of hits from convergent high-throughput studies has been suggested as a path towards finding targets with increased degrees of confidence (Rhodes et al., 2002). This enrichment can also be measured statistically using the hypergeometric test for how much the overlap across the two sets exceeds the null (Birmingham et al., 2009; R.A. Fisher, 1925; T. Nguyen et al., 2015). To characterize the statistical enrichments and the number of shared hits between the three studies of the LPS response, I applied the hypergeometric test as well as determining the degree of overlap. The enrichment of hits across studies were all of statistical significance in the hits from the positive regulator steps, however, the absolute number of shared hits was still quite limited as compared to the size of the screened gene sets (Figure 3.5A). The enrichment of negative regulator hits across studies did not meet significance in the two THP1 to RAW G9 study comparisons (Figure 3.5B). This result, however, was not as surprising since the assays used for the three studies employed close to saturating doses of the LPS ligand (N. Li et al., 2015), leaving limited scope to observe an increased response. Despite crossing thresholds of significance, however, the paucity of shared hits across these related studies is further highlighted when looking at the <10 hits shared between all three screens (Figure 3.6). These comparisons suggest that the hits

selected only by highest rank from the three studies contain multiple false positives and false negatives.

Enrichment of TLR pathway genes in high scoring hits from the THP1 TNF-α study

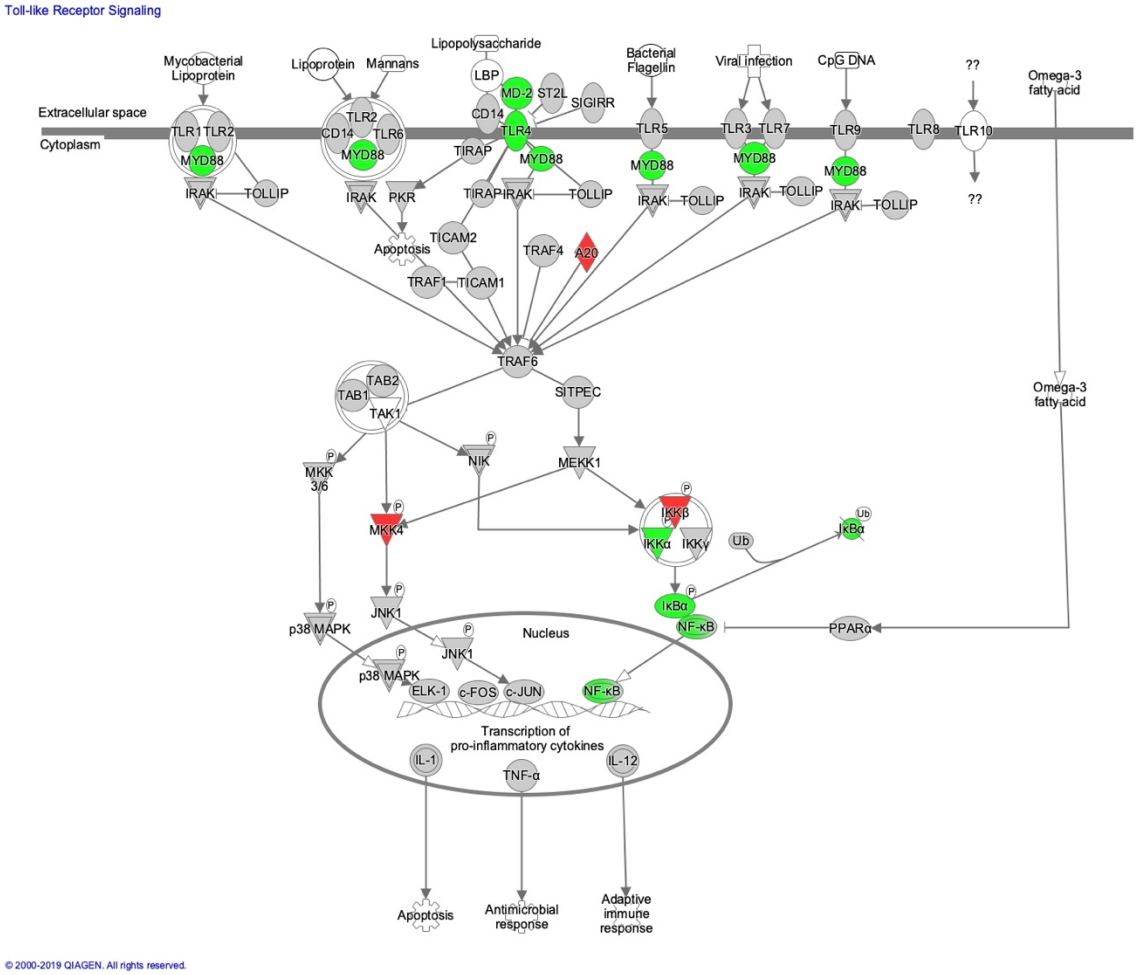


Putative Positive Regulators Putative Negative Regulators

Figure 3.2: **Enrichment of canonical toll-like Receptor pathways genes in hits from THP1 TNF-α genome-wide screen.**

Overlaying highest scoring positive (green) and negative (red) regulators from the THP1-dual luciferase study over the canonical pathway map curated by Ingenuity Pathway Analysis (IPA, Qiagen)

Enrichment of TLR pathway genes in high scoring hits from the Raw G9 NF- κB study

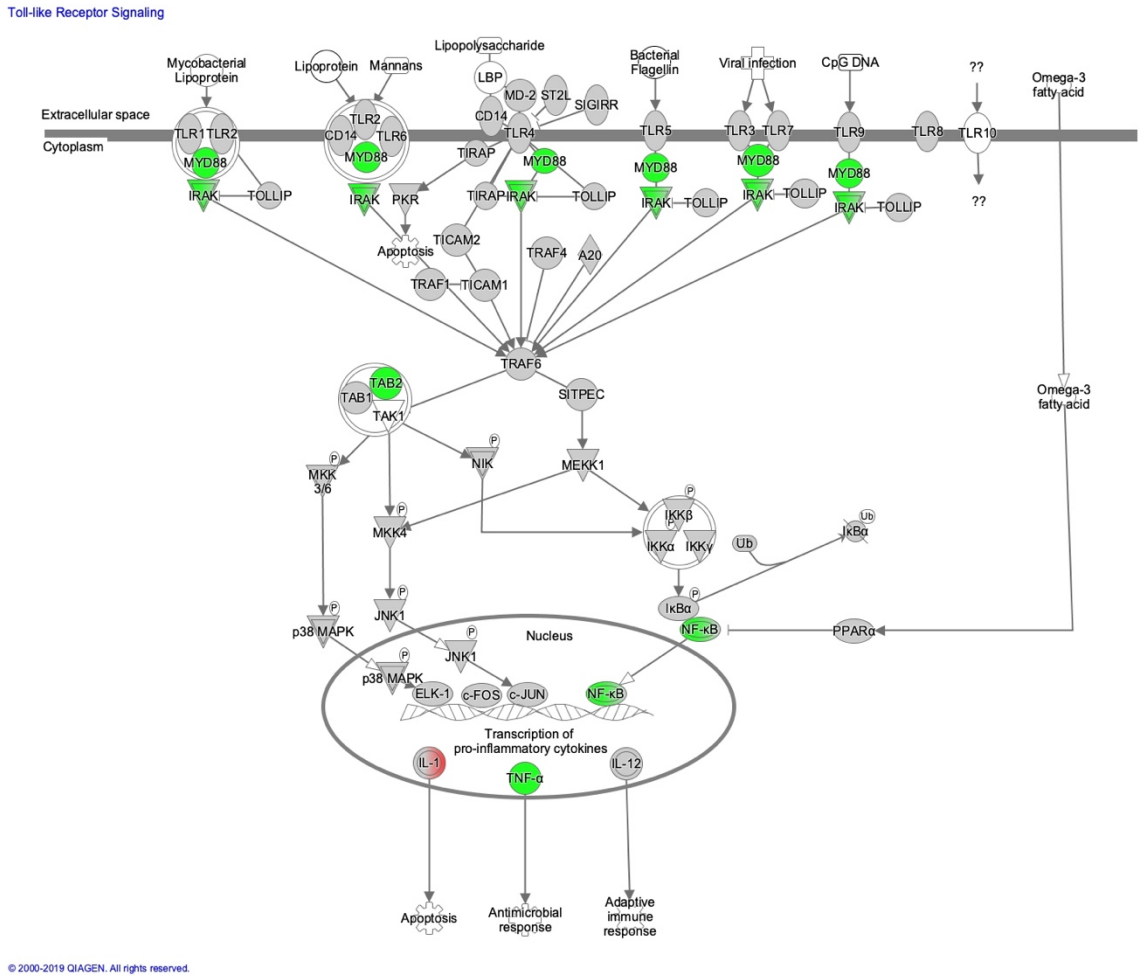


Putative Positive Regulators Putative Negative Regulators

Figure 3.3: **Enrichment of canonical toll-like Receptor pathways genes in hits from Raw G9 NF- κB genome-wide screen.**

Overlaying highest scoring positive (green) and negative (red) regulators from the RAW G9 – GFP tagged NF-κB study over the canonical pathway map curated by Ingenuity Pathway Analysis (IPA, Qiagen)

Enrichment of TLR pathway genes in high scoring hits from the Raw G9 TNF-α study



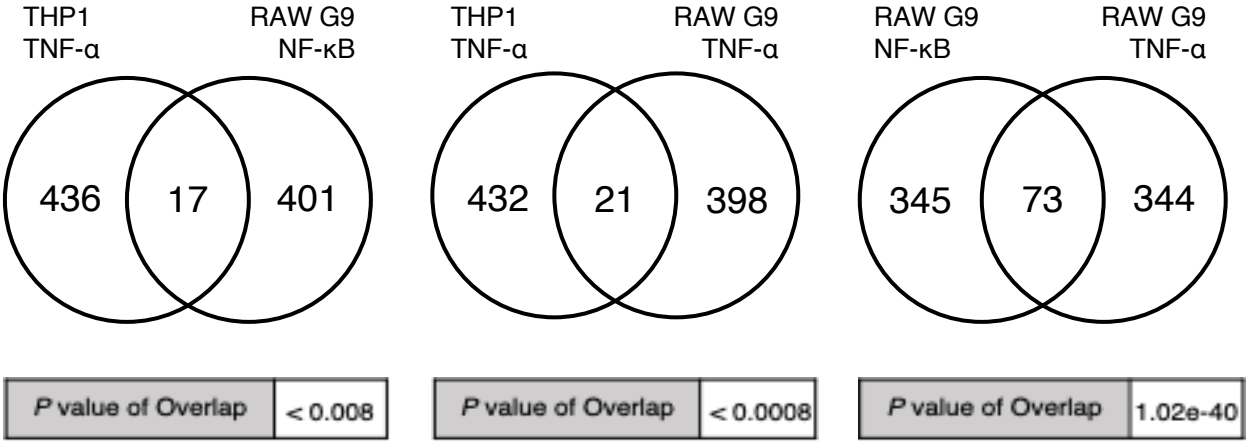
Putative Positive Regulators Putative Negative Regulators

Figure 3.4: **Enrichment of canonical toll-like Receptor pathways genes in hits from Raw G9 TNF-α genome-wide screen.**

Overlaying highest scoring positive (green) and negative (red) regulators from the RAW G9 – mCherry tagged TNF-α study over the canonical pathway map curated by Ingenuity Pathway Analysis (IPA, Qiagen)

Overlap across high scoring hits from LPS siRNA screens

A. Putative Positive Regulators



B. Putative Negative Regulators

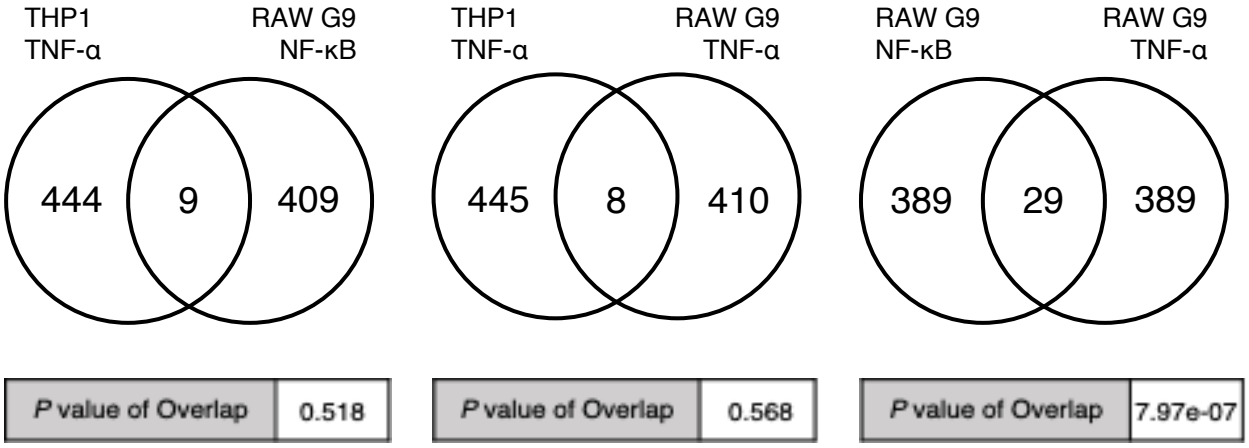
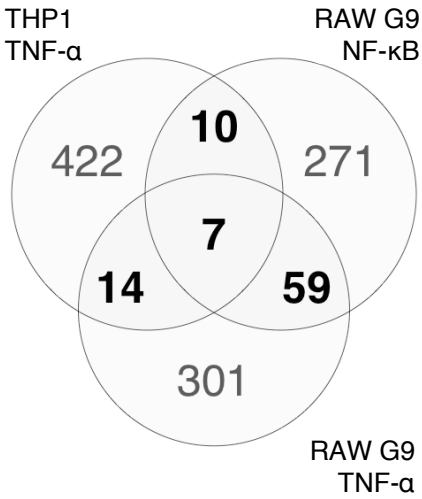


Figure 3.5: **Enrichment and significance of overlap across high scoring hits from three siRNA screens of the macrophage response to LPS.**

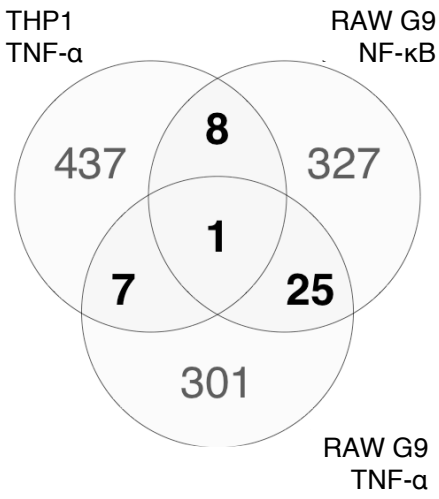
(A) 3 pairwise comparisons of the overlap between high scoring positive regulators for the THP1-dual luciferase study and RAW G9 – GFP tagged NF-κB study (left), pairwise comparison of high scoring positive regulators for the THP1-dual luciferase study and the RAW G9 – mCherry tagged TNF-α study (center), and pairwise comparison of high scoring positive regulators for the RAW G9 – GFP tagged NF-κB study and the RAW G9 – mCherry tagged TNF-α study (right). (B) Similar analysis as in A using high scoring negative regulators from the three studies.

Overlap across high scoring hits from three LPS siRNA screens

Overlap of high scoring positive regulators



Overlap of high scoring negative regulators



Gene symbols of shared high scoring positive regulators

SPI1
COPB1
NXF1
SPAG1
KAT7
ARCN1
MYD88

Gene symbols of shared high scoring negative regulators

RBMX

Figure 3.6: **Shared hits across high scoring hits from three siRNA screens of the macrophage response to LPS.**

Venn Diagrams of the shared enrichments across the three siRNA screens for selected positive regulators (left) and selected negative regulators (right). Gene symbols of the shared regulators are tabulated below each figure. (High scoring regulators from the two mouse screens were converted to their human orthologue gene symbols for comparison.)

3.5 Overlap with hits from CRISPR/Cas9-based screen of the LPS response in dendritic cells is similarly limited

A genome-scale study of the response to LPS in differentiated hematopoietic cells was also published by Parnas *et al*, 2015 (Parnas et al., 2015). While critical differences exist in the design and assay of their study and the three LPS studies I analyzed, such as cell type (Parnas and colleagues used primary bone marrow derived dendritic cells from mice) and gene perturbation methods (Parnas and colleagues used CRISPR/Cas9-based gene editing), the commonality of the TLR4 gene transcription response across these cell types would suggest that a significant enrichment of hits across these four studies would be expected. The data provided by Parnas and colleagues had a normal distribution of Z-scores and a confidence measure of FDR that aligned with a ranking of the hits by Z-score. I used a similar approach for hit selection as with the three siRNA studies, assigning a cutoff for the top 2.5% of hits (Figure 3.7A; unlike the siRNA studies, the CRIPSR study by Parnas and colleagues had transformed the readouts to have a positive Z-score for putative positive regulators from the screen). Similar to the limited enrichment between the three siRNA studies, however, concordance between each individual siRNA screen and the CRISPR/Cas9 screen was likewise limited to a minimal fraction of hits. Two of the three comparisons crossed a threshold of statistical significance with the number of shared hits ranging from 13 to just over 20 (Figure 3.7B). This comparison further highlights the limited agreement found in hit selection by different studies.

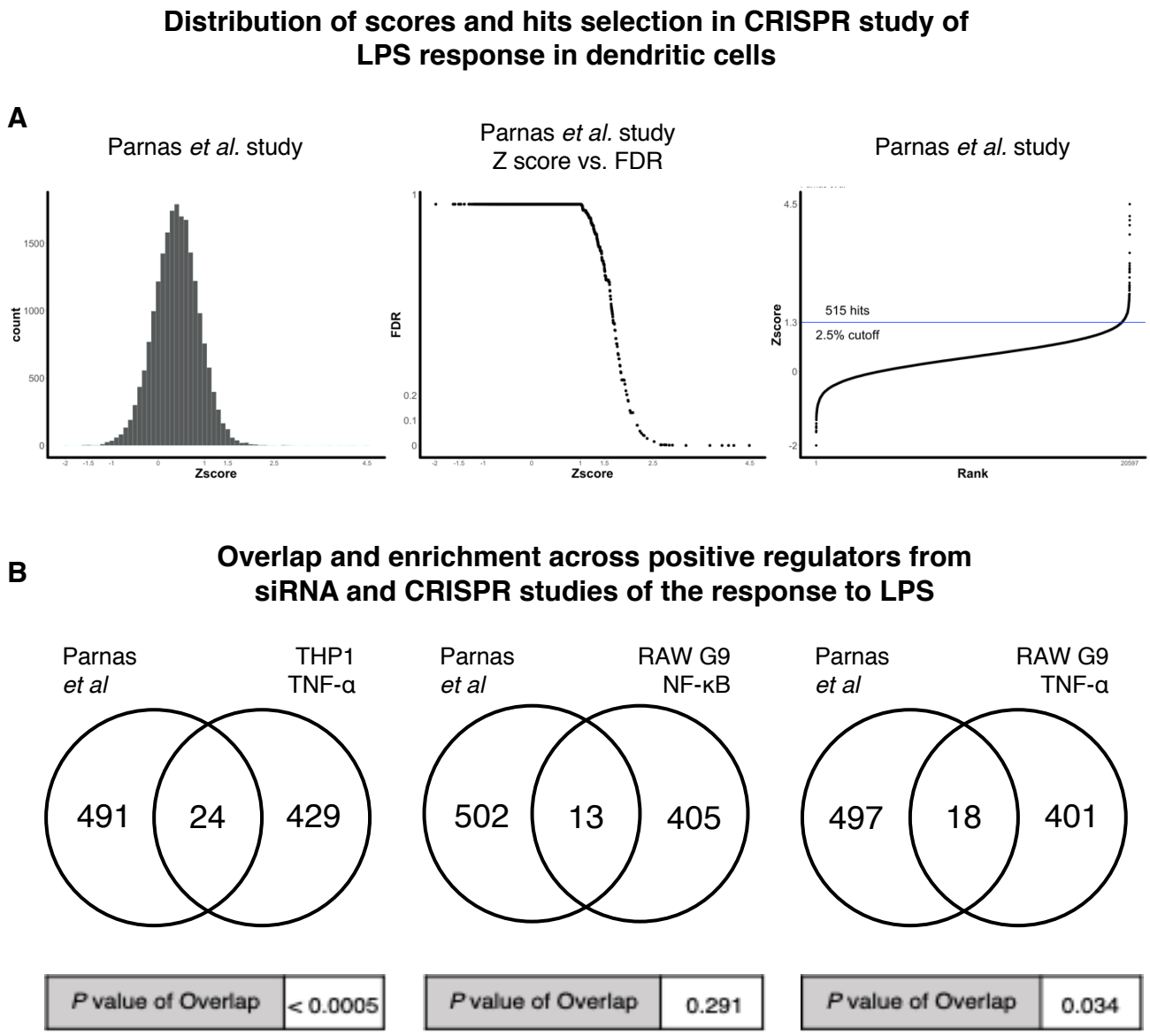


Figure 3.7: **Enrichment and significance of overlap across high scoring hits from three siRNA screens of the macrophage response to LPS and high scoring hits from the CRISPR/Cas9 screen by Parnas *et al*.**

(A) Distribution of Z scores (left) and the Z-scores plotted across assigned FDR values (center) of the data from Parnas and colleagues. Assigning a cutoff for hits in the top scoring 2.5% (right) results in a cutoff of 1.3 and 515 hits. (B) 3 pairwise comparisons of the overlap between high scoring positive regulators for the CRISPR study by Parnas *et al.* and the siRNA THP1-dual luciferase study (left), pairwise comparison of high scoring positive regulators for the for the CRISPR study by Parnas *et al.* and siRNA study using RAW G9 – GFP tagged NF-κB reporter (center), and pairwise comparison of high scoring positive regulators for the for the CRISPR study by Parnas *et al.* and from the siRNA study using the RAW G9 – mCherry tagged TNF-α reporter cell line (right).

3.6 Heterogeneity of high scoring hits across parallel omic-scale studies suggest alternative approaches are required for hit selection

The lack of concordance across the high-scoring hits from the four LPS-response studies align with a broad trend of limited overlap found in comparative omic analyses. Limited overlap of reported hits from high-throughput studies have been associated with screens of essential factors for influenza infection and early infection of HIV (as I reviewed in chapter 1 section 1.6.1). Meta-analysis has suggested that the limited overlap is driven by the approaches to hit selection employed by the different research groups (Bushman et al., 2009; Hao et al., 2013). In addition to validating our findings in the comparative analysis of the LPS studies, the previously reported HIV screens provide datasets that can be studied for how alternative hit selection methods improve enrichment and error correction.

I used the three studies of HIV Dependency Factors (HDFs) described in the introduction (section 1.6.3), the Brass *et al.* study (Brass et al., 2008), König *et al.* study (König et al., 2008), and the Zhou *et al.* study (Zhou et al., 2008). Applying normalization to the raw readout scores and appropriate cell viability corrections (described in detail in section 2.3.3) all three studies[1] generated normalized distributions of hits and rankings (Figure 3.8A). Reported hits from the three studies of HDFs were selected by combinations of subsequent analysis and validation approaches that varied from study to study. The similarities and differences in the design and hit selection of the three studies are tabulated in Table 3.2.

To see how the post-validation hits from each study related to the highest scoring hits, I mapped the normalized Z scores and confidence scores for each study and highlighted the reported hits. The final lists of hits reported by each study differed substantially from the highest scoring hits (Figure 3.8B). The sets of highest scoring hits and reported hits from the

[1] Brass *et al.* and Zhou *et al.* performed two studies one at 48 hours post infection and another at a later timepoint. For comparative purposes I'm only comparing the first study from Brass and Zhou to the study of König *et al.* to focus on the candidates regulating early infection.

three HDF studies provide a testing set for how the measures of overlap are (and are not) improved by more circuitous and multi-step approaches, as compared to the direct hit selection process by screen score alone.

		Zhou *et al.*	Brass *et al.*	König *et al.*
Experimental Conditions And Design	Cell Type	HeLa P4/R5 Cells	HeLa-derived TZM-bl Cells	293T Cells
	Treatment	HXB2 HIV- 1	HIV IIIB	VSV-G pseudotyped HIV-1 reporter virus encoding luciferase
	Readout 1	Tat activation of expression of the β-Gal reporter	p24 (product of gag gene)	HIV-1 Vector encoded luciferase
	Time point: Readout 1	48h	48h	24h
	Readout 2	Tat activation of expression of the β-Gal reporter	β-Gal (Tat dependent)	MuLV and AAV
	Time point: Readout 2	96h	72h	24h
Hit Selection, Bioinformatics, and Secondary Screening	Cell Viability Correction	Decrease of cell viability by 2 SDs or more	Decrease of cell viability by 2 SDs or more	Cell toxicity screen
	Z score cutoff	2 SSMD relative to the negative control	2 SDs greater than the plate mean	2 siRNAs with ≥45% reduction in HIV infectivity
	Bioinformatics Used in Hit Selection	In silico screening for expression in activated T cells and Macrophages	None	"evidence score" based on functional, biochemical, and transcriptional data. Yeast to hybrid protein interaction database, NCBI HIV-1 Protein Interaction Database, MCODE, Ontogeny-based pattern identification algorithm
	Secondary Screening	Rescreening by independent siRNAS	Rescreening of pooled siRNAs in single siRNA assay	Rescreening of pooled siRNAs in single siRNA assay

Table 3.2: **Design and hit Selection methods for three siRNA studies of early HIV dependency factors by Zhou *et al.*, Brass *et al.*, and König *et al.***

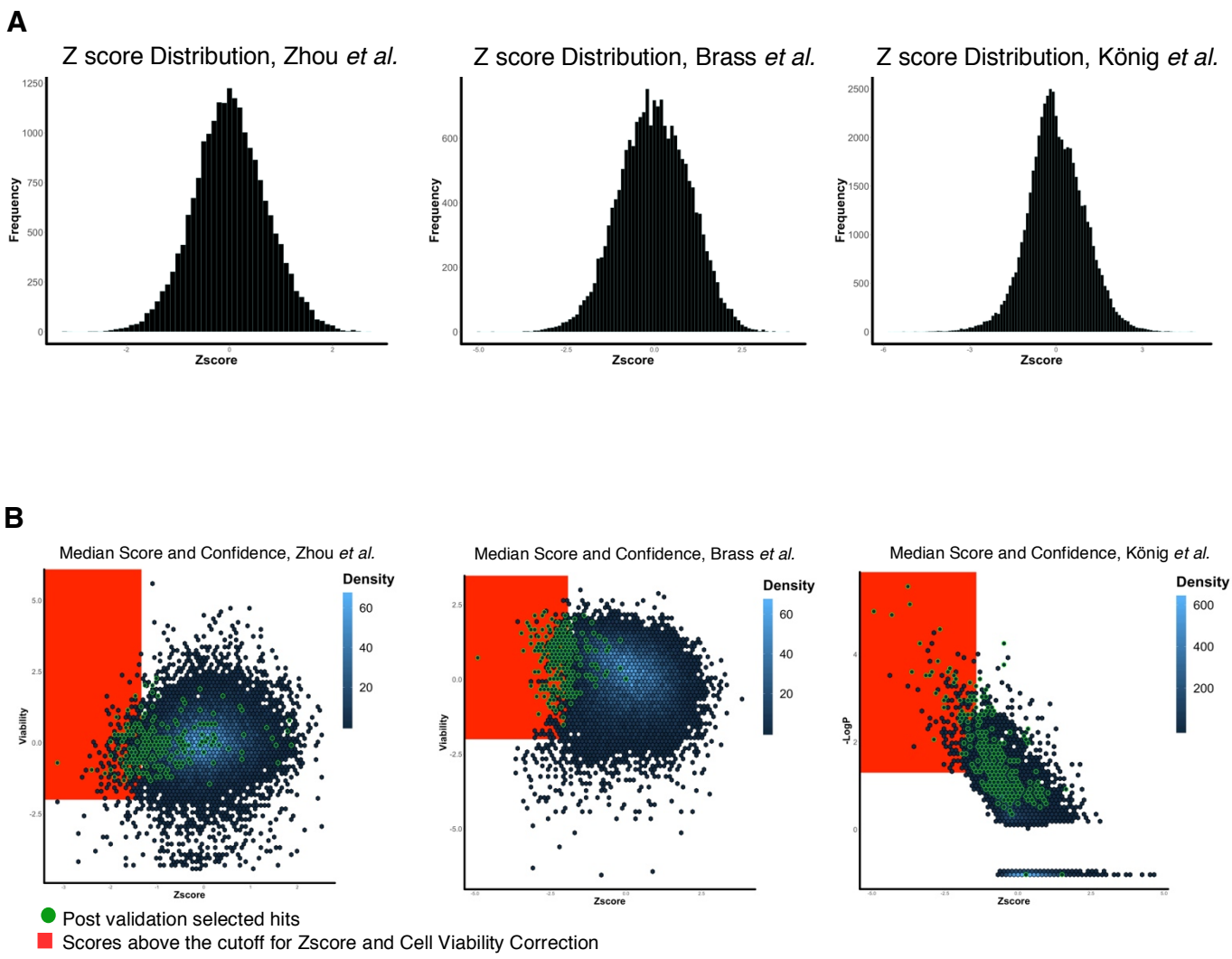


Figure 3.8: **High Scoring and Post-Validation Selected Hits from Three Studies of HIV Dependency Factors.**

(A) Distribution of normalized Z scores for siRNAs from the study by Zhou *et al.* (left) Brass *et al.* (center) and König *et al.* (right) B) Identifying the candidates selected by high score and those selected as post validation hits by the three studies. The post validation hits are in green and the area on the graphs where hits fall above the high scoring cutoff is highlighted in red. The density of hits at each point is represented by the blue color scale. Zhou *et al.* (left) and Brass *et al.* (center) used a cell viability count that was normalized to a Zscore. König *et al.* (right) used a p Value that was a combined "evidence score" that was plotted as a negative log p Value.

3.7 Statistical enrichment, but not shared hits, is narrowly improved by commonly applied bioinformatic approaches to hit selection

Utilizing the two hit selection sets (High scoring and Post-validation) identified in the previous section, I did a comparative analysis of statistical and shared enrichment across the three studies. When measuring the enrichment of Post-validation hits, all three comparisons crossed standard thresholds of statistical enrichment, while only one of the comparisons of hits selected only by high scores met the significance threshold (Figure 3.9A). This change reflects the significant improvement achievable by the bioinformatic and secondary validation methods used by the three studies beyond the high scoring results. When counting the absolute number of shared hits across the three screens, however, the numbers are very low and there is no discernible improvement between the high scoring hit sets to the post validation selected hit sets (Figure 3.10A-B). To understand further how this dichotomy emerges, I did a comparative analysis of the two critical measurements of overlap that drive the calculation of statistically significant enrichment (described in detail in the introduction section 1.6.2).

The number of shared hits between each comparison of two studies (corresponding to the measure of *m* in the hypergeometric testing equation (Figure 1.7)) remained nearly consistent between the sets of hits selected by high scores and the sets of hits selected by subsequent and secondary validation (Figure 3.9B). When calculating the relative size of the selected hit sets, however, there is a clear reduction in the set size (n_1 and/or n_2) when moving between the sets of hits selected by high scores versus the sets of hits selected by the different analysis and validation pipelines (Figure 3.9C). These findings show that the increase in statistical enrichment gained by the post validation hits are largely driven by reducing the number of hits selected and eliminating hits that have lower likelihoods of being shared by comparative studies. These results further suggest that these methods are less effective when it comes to increasing the absolute number of hits that are identified across multiple studies. The

added bioinformatic and follow up analyses used for the hit selection of the reported hits from the three studies of HIV HDFs substantially reduced the number of false positives in the hit selection sets. The limited increase in overlap, however, shows that these approaches did not reduce the false negative rate as lower scoring hits missing the high score cutoff were not corrected for and considered as hits by the secondary analysis steps.

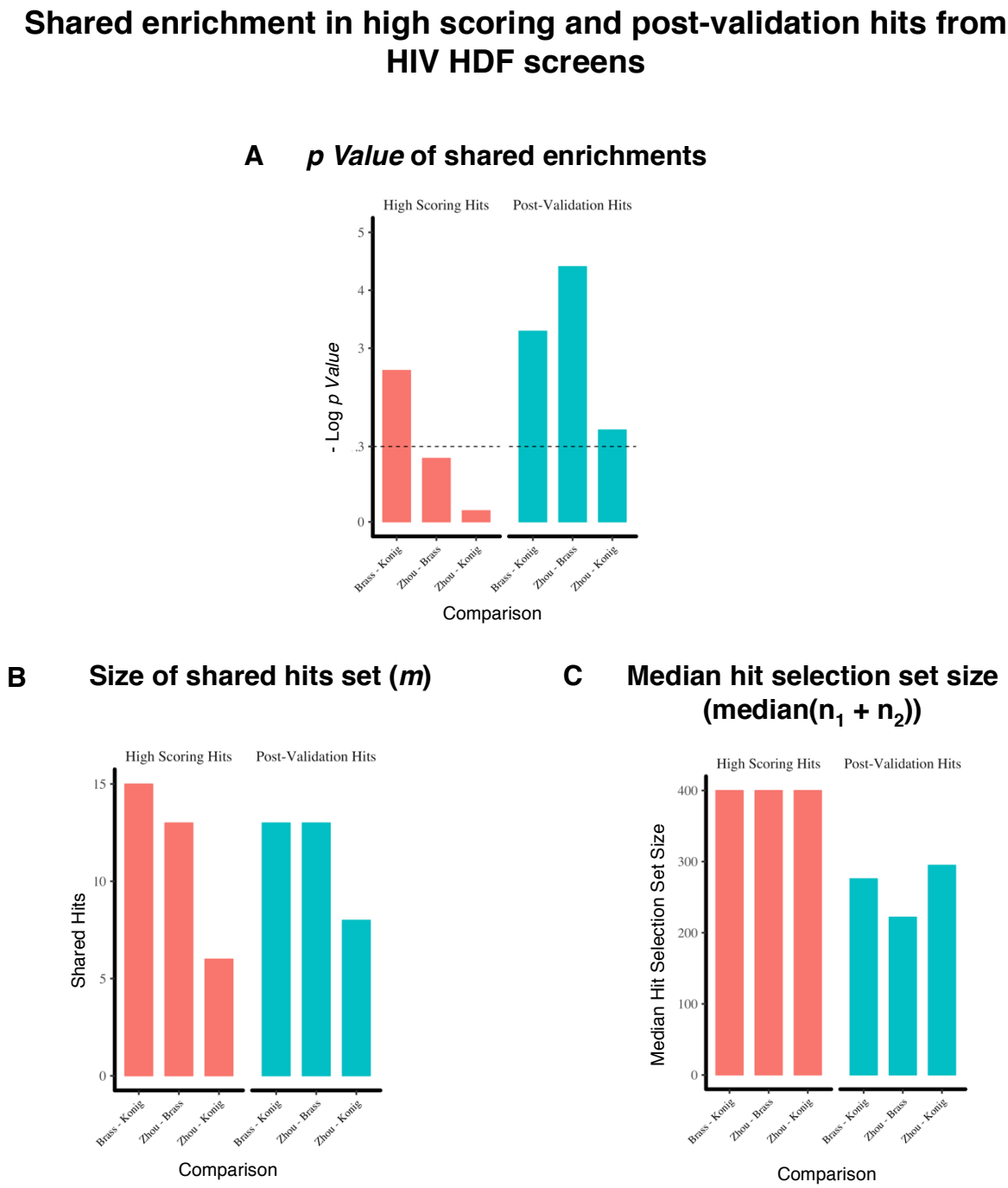


Figure 3.9: **Measurements of shared enrichment and overlap in high scoring and post validation selected hits from three siRNA studies of HDFs.**

(A) Negative Log p Values of the statistical enrichment between the hit sets of the three siRNA HDF studies for, both, hits selected by high score cutoff (red) and hits selected by analysis and secondary screening (blue). The point at the y axis that corresponds to $p = 0.05$ is indicated with a dashed line across the graph. (B) Number of shared hits in two screen comparisons of the three siRNA studies of HDFs for hits selected by high score cutoff (red) and hits selected by analysis and secondary screening (blue). (C) Median size of the hit selection sets for each two way comparison of the three siRNA studies of HDFs for hits selected by high score cutoff (red) and hits selected by analysis and secondary screening (blue).

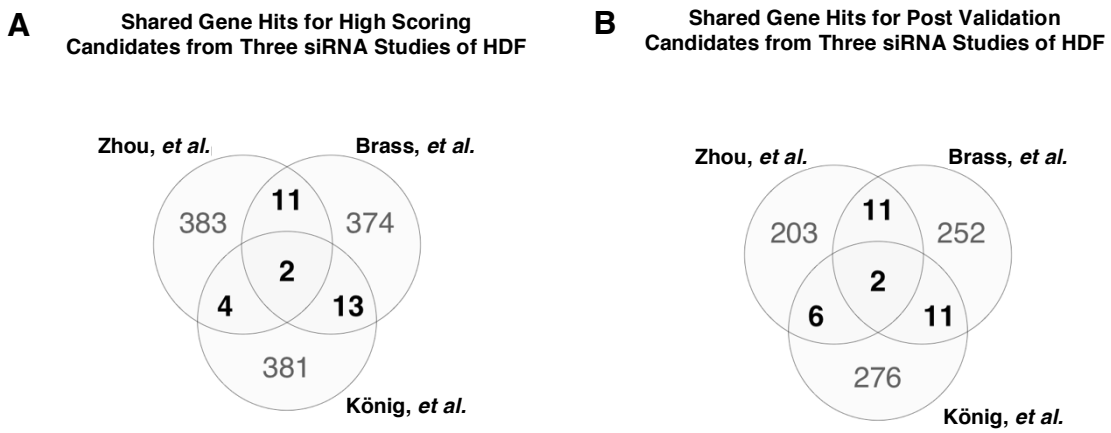


Figure 3.10: **Overlap in high scoring and post validation selected hits from three siRNA studies of HDFs.**

Venn diagram of shared and unshared hits for the three studies of HIV HDFs selected by highest score (A) and from the post validation sets (b).

3.8 Summary

The genome-scale screens for putative regulators of the TLR4 signaling pathway in macrophages provides a platform for the robust characterization of novel regulatory targets that go beyond the canonical and previously characterized members of the pathway. This approach is limited by the challenge of devising a hit selection method that overcomes the variation and noise intrinsic to studies with omic scale measurements. This challenge is reflected in the low level of commonality found across many comparative genome-scale studies and is similarly mirrored in my analysis of the described LPS response screens in differentiated hematopoietic cells. Devising a method for hit selection that is better at identifying shared enrichments has been a critical obstacle in genome-scale analyses, as a degree of expected concordance across studies is a critical metric to establish confidence in methodology and execution of high-throughput studies.

My analysis of hit selection in three published studies of early HIV infection shows that the current practice of cutoff setting and supplementation through divergent enrichment analysis methods improves the statistical significance of screen overlap. These same analytical approaches, however, had no impact on the total number of hits identified across more than one study. The number of candidates identified by more than one genome-scale study remained nearly steady whether using a direct normalization and high score cutoff approach or if using the analysis and secondary validation selected by the three research groups. This suggested that robust hit selection from the three studies of the LPS response in macrophages would require the development of a new approach to increase the number of shared enrichments, while also reducing the number of candidates unlikely to appear in repeated assays. In the next chapter I again use the three studies of HIV HDFs as a testing dataset to assess alternative approaches to hit selection from high-throughput assays.

4 Developing a TRIAGE approach for hit selection from high-throughput data

4.1 Introduction

In the previous chapter I have shown that bioinformatic analysis approaches could improve screen concordance in the three studies of HIV HDFs. I have also shown that the improvement was largely concentrated in one metric, significance of overlap, and that it was driven by the strong removal of false positives by the added analyses. A different metric of hit selection confidence, number of shared hits, that is driven by low false negative rates, was not improved by the employed hit prioritization methods. The bias towards false positive correction methods is to be expected with secondary follow up studies. Secondary low-throughput follow up assays are designed to remove false positive hits and are critical to gaining confidence in the results from a screen. This approach, however, lacks a mechanism to correct for false-negatives. As I have argued in the introduction, many results and insights from high-throughput studies are compromised by high false negative error rates (citing examples in the literature from studies in influenza (Hao et al., 2013), HIV (Bushman et al., 2009), and β-catenin-active cancers (Rosenbluh et al., 2016)). Simply correcting false-negative rates by increasing the number of hits selected for follow up (such as lowering the score cutoff or selecting additional hits for secondary analysis) is costly and inefficient. There is a pressing need for hit selection methods that can correct for the false negative rates while also generating list of hits that can be reasonably validated in lower-throughput follow up experiments.

In this chapter I propose an alternative method for how to assign hits vs. non hits from normalized readout scores (section 4.2). I show, using the HDF studies as an example, that pathway analysis and network analysis improve hit selection in different ways (sections 4.5-4.8). I propose and validate a framework for integrating these approaches to get the optimal combined result from the different methods (sections 4.9-4.12). I name the framework Throughput Ranking by Iterative Analysis of Genomic Enrichment (TRIAGE). I then show

how TRIAGE analysis can also be used to reanalyze published studies using the previously validated hits (section 4.13).

4.2 Segmentation of data by degrees of confidence is an alternative to the binary hit vs. non-hit approach

The single cutoff approach bifurcates a dataset into hits and non-hits. This approach requires an irreversible compromise between false positive and false negative correction right at the outset of downstream data analysis. As an alternative, I set out to use a dual cutoff approach in hit selection. In the dual cutoff approach two cutoffs are selected, a stringent cutoff that assigns the score for what is considered a "high confidence" hit, and a more lenient cutoff score. Candidates whose scores fall between the lenient cutoff and stringent cutoff are considered "medium confidence" hits. As a result, you get a three-tiered data set of high confidence hits, medium confidence hits, and low confidence/non-hits (Figure 4.1A). This approach can be used for data with a single readout (such as Z score or fold change) or for data with multiple readouts (such as Z score and cell viability, fold change and *p* value, etc.). Datasets with more than one readout can be segmented by assigning dual cutoffs to all readouts and then assigning as high confidence hits those that meet the stringent cutoff in two or more of the readouts and medium confidence as those that meet the stringent cutoff in only one readout. An alternative approach for datasets with more than one readout is to assign a single cutoff to one readout and a dual cutoff to another readout and require that all medium confidence and high confidence hits meet the criteria of the first cutoff (such as a critical cell viability metric or a similar confidence measure).

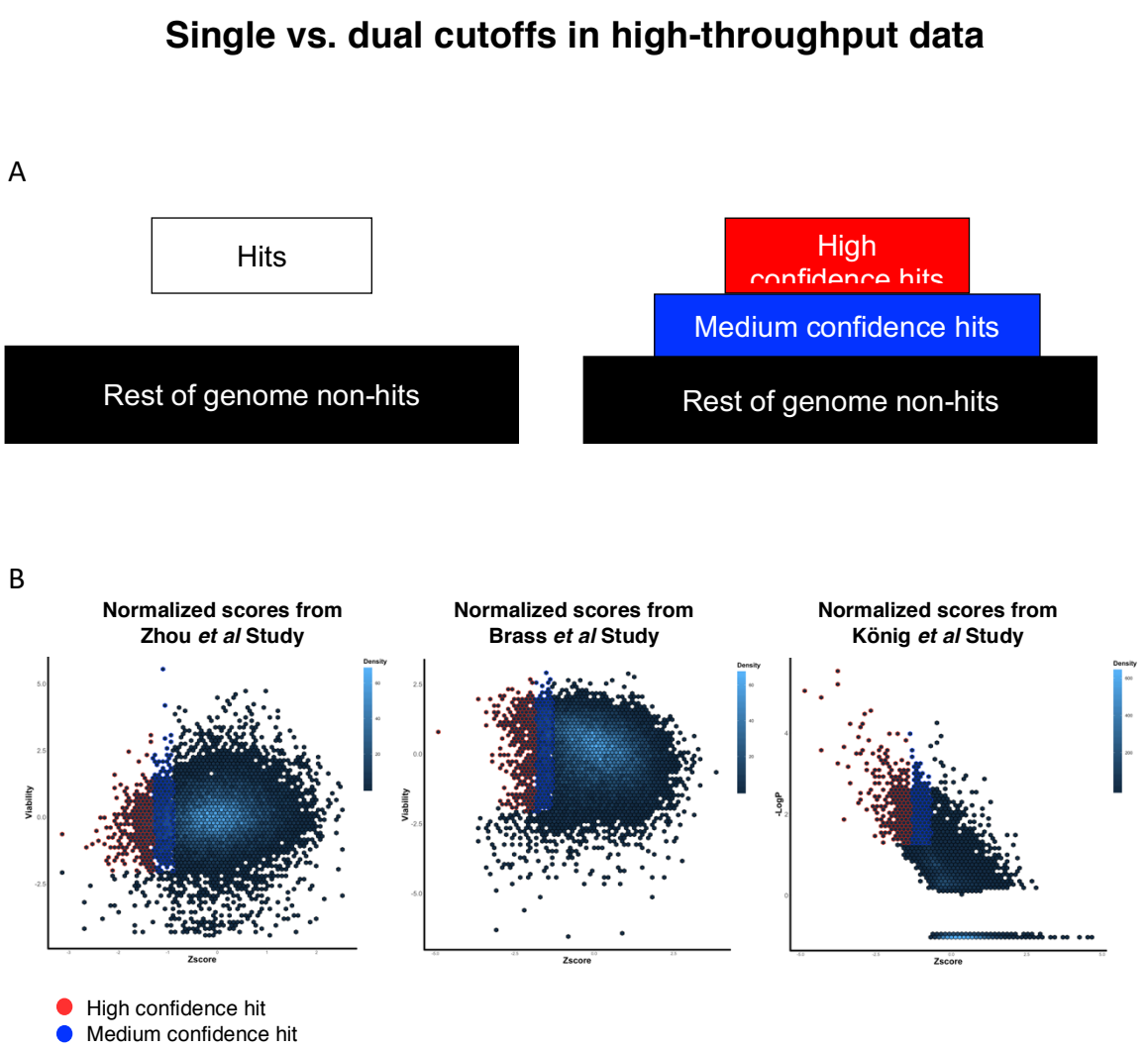


Figure 4.1: **Segmenting data from high-throughput studies into three degrees of confidence.**

(A) A schematic of the two data tiering approaches in high-throughput hit selection. A single cutoff approach (left) assigns a set of gene as hits and the rest as non-hits versus a dual cutoff approach (right) assigns a set of hits as high confidence, a secondary set of hits as medium confidence, and the rest as low confidence/non hits. (B) Scores from three genome-wide studies of HDF. Normalized scores are plotted on the x-axis and secondary scores that were considered (such as cell viability and assigned p-values) are on the y-axis. Genes with both scores above the cutoff are in red and genes with Z scores above the secondary cutoff are in blue. The density of hits at each point is represented by the blue color scale.

As an example of the three-tier data approach, I segmented the data from the three studies of HDFs introduced in chapter 3. The readouts of all three screens were normalized by Z score and each gene was assigned the median score of all its siRNAs. I then plotted those scores against the secondary scores used by the three studies (Zhou and Brass used cell viability readouts while König used a compiled "evidence score" *p* value). I assigned a single cutoff to the secondary scores of the three screens (as described in their respective publications; Zhou and Brass used a cutoff of -2 standard deviation for cell viability, König used a combination of scores and prioritized those with evidence scores of 0.05 or less). I then assigned dual cutoffs to the Z scores of the three screens. To make the set sizes comparable I assigned the genes with the 400 highest scores as high confidence hits. The gene candidates with the next 1000 highest scores (that were not included in the top 400 and also were above the threshold of the secondary readout cutoff) I assigned as medium confidence. The rest of the gene candidates were assigned as non-hits (Figure 4.1B).

4.3 Hit selection and overlap in three genome-wide studies of HDFs

In the previous chapter I have shown that the three genome-wide siRNA studies of HDFs demonstrate the gap between hits selected by normalized scores and hits selected by the respective research groups. These two hit selection approaches can be summarized as hit selection based on highest scoring hits versus user guided prioritization. The latter approach (as outlined in Table 3.2 in the previous chapter) is guided by a combination of different secondary analysis approaches that are applied to the highest scoring hits in different ways by different investigators. These supplementary analysis approaches include a combination of pathway analysis and network analysis applied to the selected hits, as well as adding in or removing hits based on user preferences. How these different analysis approaches are combined into a single set of selected hits also differs from study to study and is dependent on

the prioritization of the authors (Figure 4.2A). This approach differs from the direct and unbiased hit selection method of selecting the highest scoring candidates as hits (Figure 4.2D). These two approaches represent the two extremes of hit selection from high-throughput data, a direct and unbiased approach or a user guided combination of supplementary analysis approaches and follow up.

As I have shown in the previous chapter for the three studies of HDFs, the user guided prioritization approach for reported post validation hits from each study only improves hit selection in one measure (*p* value of shared enrichment) while having only a marginal impact on the number of shared hits (Figure 4.2B-C and 4.2E-F). For this chapter I will be using these two analysis approaches as examples to compare against alternative methods. I continue using the quantitative measures of how significance of overlap and number of shared hits between the three studies are affected as assessments of false positive and false negative correction, respectively.

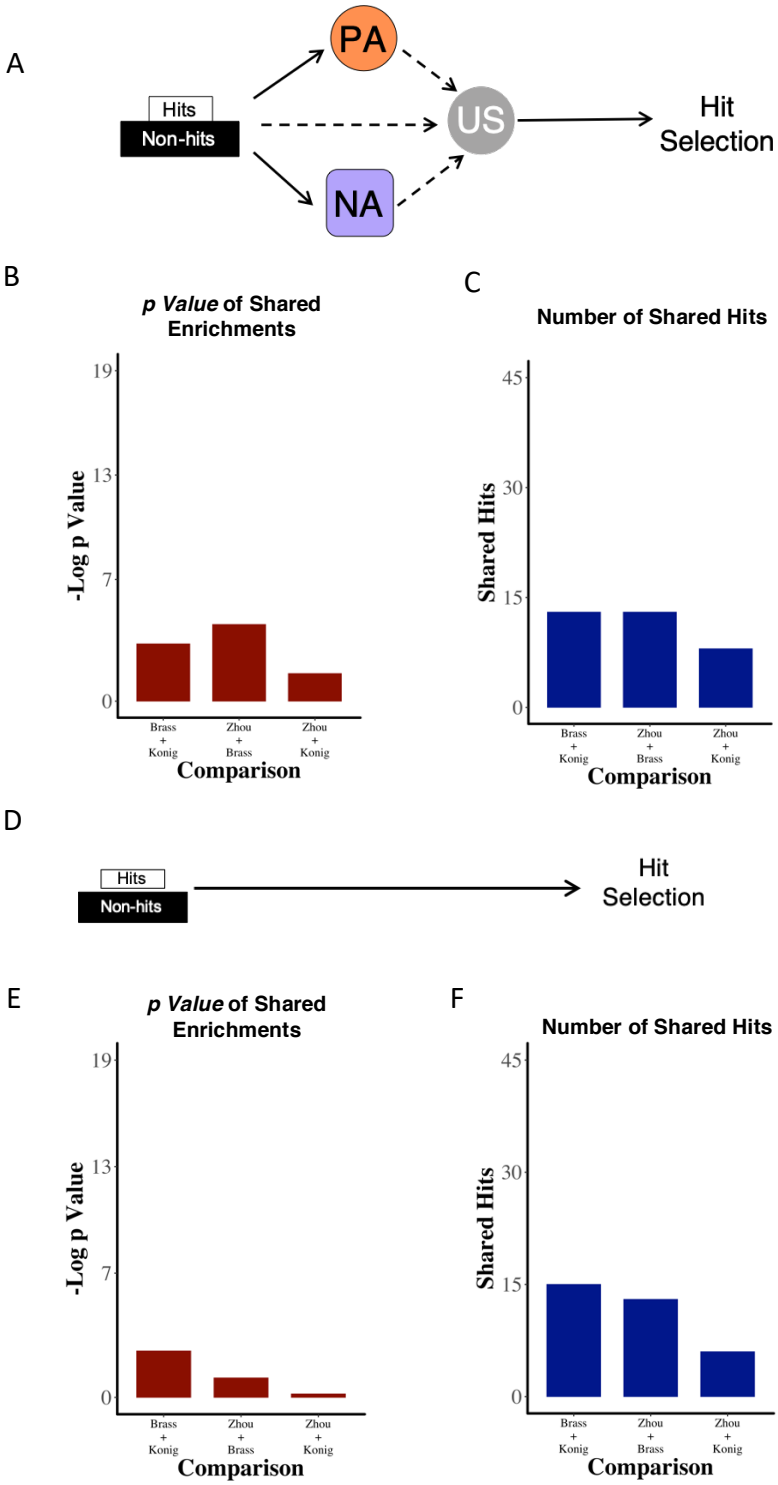


Figure 4.2: **Hit selection by user guided prioritization or highest score.**

(A) Schematic of the user guided prioritization approach for hit selection. A combination of pathway analysis (PA), network analysis (NA) and user selected hits (US) are combined by the user to form a set of primary screen hit selection that are followed up on with validation studies. (B) statistical significance of the overlap across the three studies of HDF when comparing the reported hits from the studies. (C) Number of shared hits between the reported hits from each study. (D) Schematic of the hit selection approach by highest score. (E) statistical significance of the overlap across the three studies of HDF when comparing the highest scoring hits. (C) Number of shared hits between the highest scoring hits from each study.

4.4 Random permutation testing

To evaluate how different bioinformatic approaches do or do not improve hit selection from high-throughput studies it is critical to first develop a test that can measure our confidence in these results. Bioinformatic approaches have intrinsic biases such as overrepresented groups and low resolution (Khatri et al., 2012). Given a sufficient number of hits to choose from it is possible that the biases within these analyses could lead to similar hits being selected from unrelated datasets. To increase confidence that the results represented in my findings are driven by the prioritization of biologically relevant candidates and not by the overrepresentation and annotation sets of the databases and methods used, I applied a random permutation testing approach. For each analysis and comparison that follows in this chapter I generated 1000 input files that have the same size of hits and non-hits (or high confidence hits, medium confidence hits, and non-hits where relevant) with the gene candidates assigned to different confidence groups at random. I ran each of the randomly generated inputs through the same analysis and cross-screen comparison as the non-random input. I then plotted the number of shared hits found in each run of the random input analysis. Using an empirical cumulative distribution I calculated what quantile in the distribution of random results the value from the non-random input corresponded to. (i.e. the frequency of different results in the randomly generated input was taken as a measure of how likely the results we found in our analysis was to be driven by the biases of the analysis method or by the size of our input versus being a non-random biologically relevant result.).

To test the strength of this approach, I applied the random permutation test to the comparison analysis for hits selected by high scores. As no additional analyses were applied to the selection of these hits, the random permutation test results should align with results from the hypergeometric test (one-sided Fischer's Exact Test). A comparison of the two tests found that the results were closely aligned (Figure 4.3).

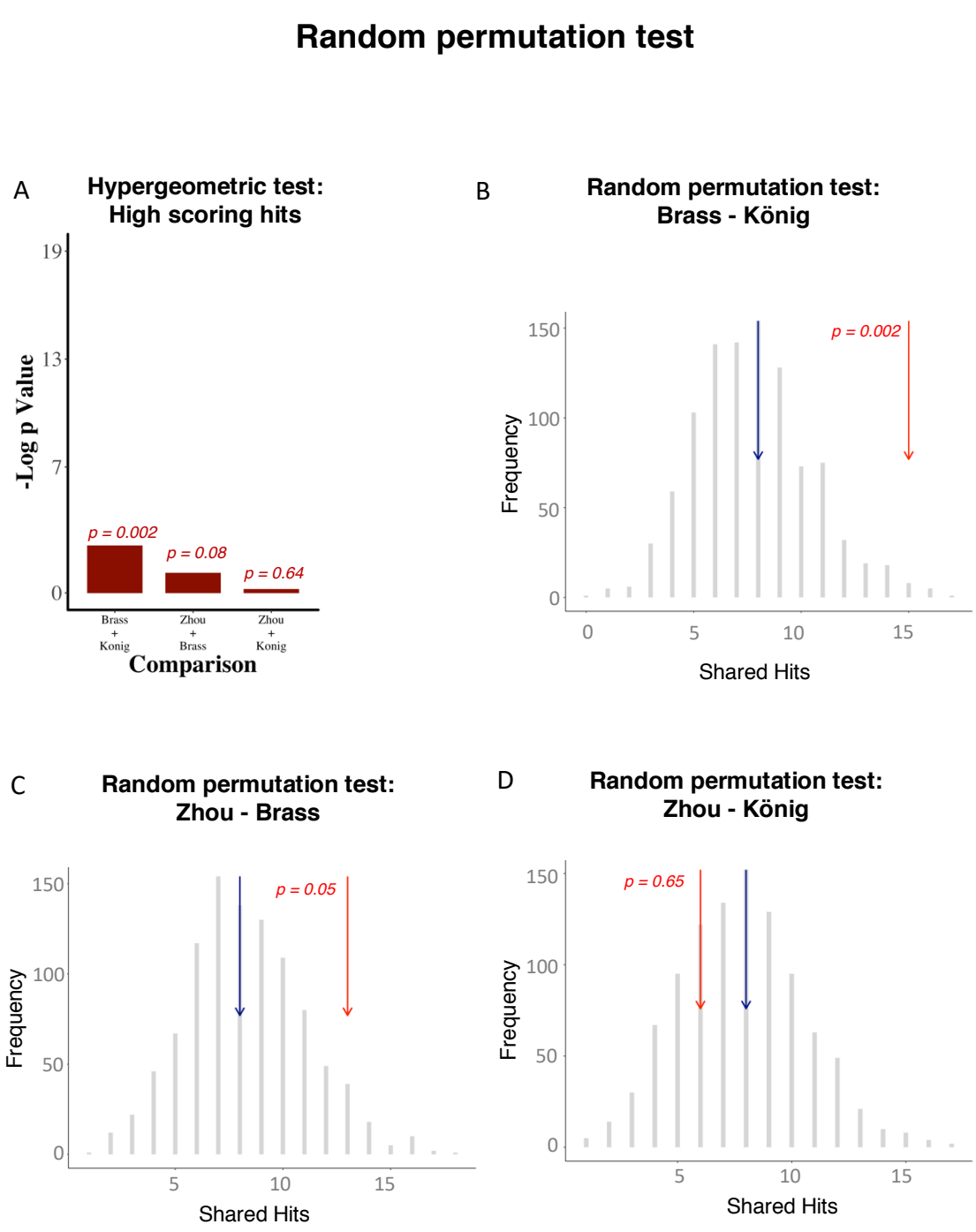


Figure 4.3: **Random permutation testing of highest score hit selection.**

(A) *p* values for shared enrichments across the studies of HDF as calculated by the hypergeometric test. (B-D) Distribution of shared hits by 1000 random permutation of equal size hits and datasets. Red arrow indicates where the number of shared hits from the non random dataset falls in the distribution. Blue arrow indicates where the median size of shared hits by random permutation falls in the distribution. Number is calculated by 1-quantile.

4.5 Hit selection by pathway analysis of high scoring hits improves statistical enrichment in some cases

To test how the application of pathway analysis to high-scoring hits improves measures of overlap in hits from HDF studies, I applied pathway analysis to the high scoring hits of the three screens. I used the pathway membership list from the KEGG database, and applied a filtering only for pathways that are related to biological processes, and removed pathways that are related to disease networks (full process described in section 2.2.1 in Material and Methods). I then applied the hypergeometric test for enrichment of each pathway. High scoring genes that were in a pathway that had an enrichment score of $p \leq 0.05$ were selected as hits, all high scoring hits not in an enriched pathway were reassigned as non-hits (Figure 4.4).

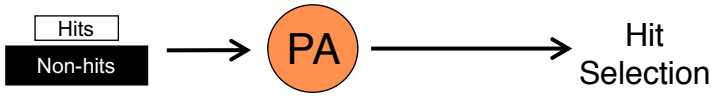


Figure 4.4: **Hit selection by pathway analysis.**

Schematic of the pathway analysis approach for hit selection. Candidates are divided by a single cutoff. Pathway analysis (PA) is applied to hits. Hits from enriched pathways are selected as final hits.

Statistical significance of overlap increased in two out of the three comparisons as compared to the significance of overlap in high scoring and post validation hits (Figure 4.5A). Number of shared hits, however, decreased in all cases as compared to the alternative analysis approaches. These results align with what would be expected from a direct pathway analysis approach. Pathway analysis is designed to filter the set of hits and remove false positives (as is reflected in the hypergeometric test results). Pathway analysis, however, does not include a false-negative correction mechanism and can in some cases increase false negative rates by biasing the results away from less annotated hits. This trade-off is reflected in the number of shared hits decreasing across all comparisons (Figure 4.5B). Pathway analysis is also uniquely sensitive to the setting of cutoffs, the noisier the dataset the less likely it is to find true enrichments. This sensitivity might explain why the improvement in enrichment significance in only observed in some cases.

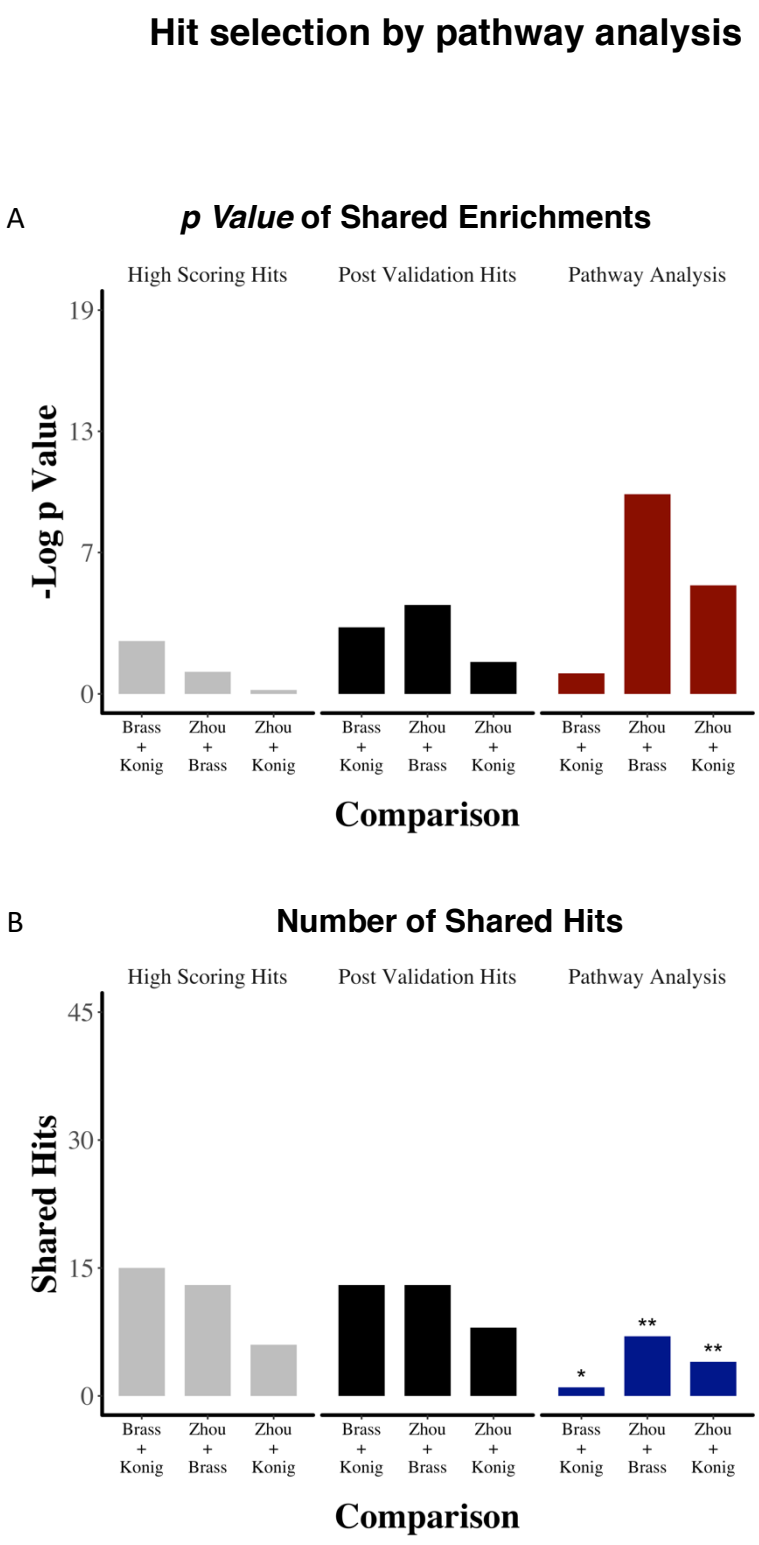


Figure 4.5: **Enrichment and overlap in hit selection by pathway analysis.**

(A) statistical significance of the overlap across the three studies of HDF when comparing hits selected by pathway analysis versus highest scoring hits and post validation hits. (B) Number of shared hits between the hits selected by pathway analysis from the three studies versus highest scoring hits and post validation hits. Random permutation test scores: ns = $p > 0.05$, * = $p \leq 0.05$, ** = $p \leq 0.01$

4.6 Hit selection by pathway analysis of three-tiered data has a stark impact on the statistical enrichment, but not commonality, of hits across studies of HDFs

As an alternative approach, I ran pathway analysis on the HDF screens using a dual cutoff method applied to three-tiered data. Following analysis of the high confidence set to first establish enriched pathways, high confidence and medium confidence hits that were members of those pathways were brought into the hit set. Hits in either the high confidence or medium confidence set that were not part of these enriched pathways

were reassigned as non-hits (Figure 4.6). The strength of this approach is that it relies on the high confidence set of hits to guide the pathway analysis and then expands the hit selection to include lower scoring hits that are part of enriched-for biological processes.

A comparison of the three studies of HDFs with hit selection by this approach shows that it strongly improves both measures of overlap (significance of enrichment and size of shared hits) in all screens as compared to the single cutoff pathway analysis approach (figure 4.5 A-B and Figure 4.7A-B). Comparison to hit selection by high scores and post validation, however, shows that the strength in pathway analysis is predominantly in the false positive correction (Figure 4.7A), and only marginally adds to the false negative correction (as measured by the commonality of hits across studies) (figure 4.7B).

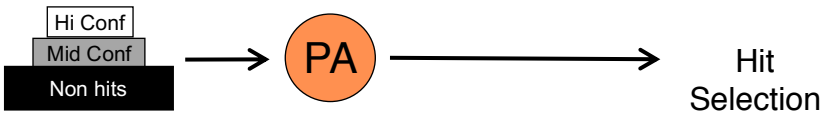


Figure 4.6: **Hit selection by pathway analysis using dual cutoffs.**

Schematic of the pathway analysis approach for hit selection from a three-tiered dataset. Candidates are assigned to three groups based on hit confidence. Pathway analysis (PA) is applied to high confidence hits. High confidence and medium confidence hits from enriched pathways are selected as final hits.

Hit selection by pathway analysis using dual cutoffs

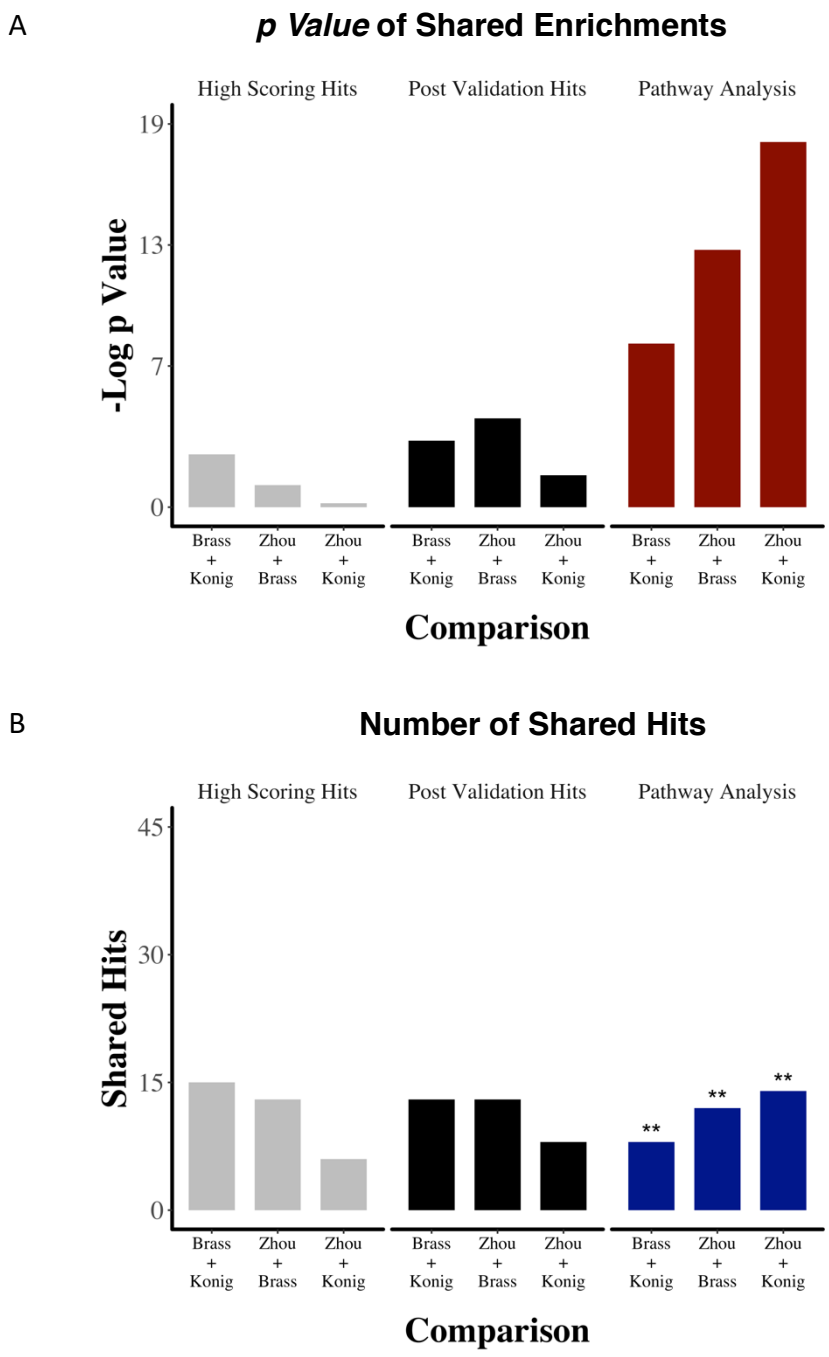


Figure 4.7: **Enrichment and overlap in hit selection by pathway analysis using dual cutoffs.**

(A) statistical significance of the overlap across the three studies of HDF when comparing hits selected by pathway analysis of three-tiered data versus highest scoring hits and post validation hits. (B) Number of shared hits between the hits selected by pathway analysis of three-tiered data from the three studies versus highest scoring hits and post validation hits. Random permutation test scores: ns = $p > 0.05$, * = $p \leq 0.05$, ** = $p \leq 0.01$

4.7 Hit selection by network analysis of three-tiered data has a strong effect on the number of shared hits, but not enrichment, between parallel studies

The other commonly applied data analysis approach to improve hit selection from high-throughput studies is network analysis. I applied network analysis to the dual cutoff datasets of the studies of HDFs. Using the protein-protein interactions curated by the STRING database and mapped back to gene IDs, I filtered the interactions for those based on published experimental evidence or curated databases (assigned "experimental" and "database" in the platform). Interactions were required to have a confidence score of medium or higher (full method described in section 2.2.2 in Materials and Methods). High confidence and medium confidence hits were entered into the network and searched for predicted interactions. The interactions were filtered to include only those between a high confidence hit and a medium confidence hit as a means to promote medium confidence hits to the high confidence set. All other medium confidence hits were assigned as non-hits (Figure 4.8).

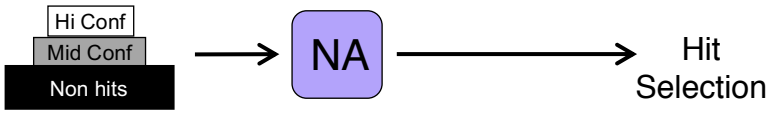


Figure 4.8: **Hit selection by network analysis.**

Schematic of the network analysis approach for hit selection. Network analysis (NA) is applied to high confidence hits. Medium confidence hits that have predicted interactions with high confidence hits are added to final hits.

Hit selection by network analysis of high confidence and medium confidence hits led to a sharp increase in the number of shared hits across studies (figure 4.9B). Significance of overlap however was not improved (figure 4.9A) reflecting the expansion of the total number of hits selected and the increase in false positive hits. The testing by random permutation also found that the number of shared hits found by network analysis alone was only above a statistical threshold of significance in two out of the three comparisons (figure 4.9B) suggesting that the 'catch-all' approach of network analysis without any false positive correction is prone to the amplification of false positives. This suggests that hit selection by network analysis is a highly sensitive approach, but needs additional correction to increase the specificity of the hit selection set.

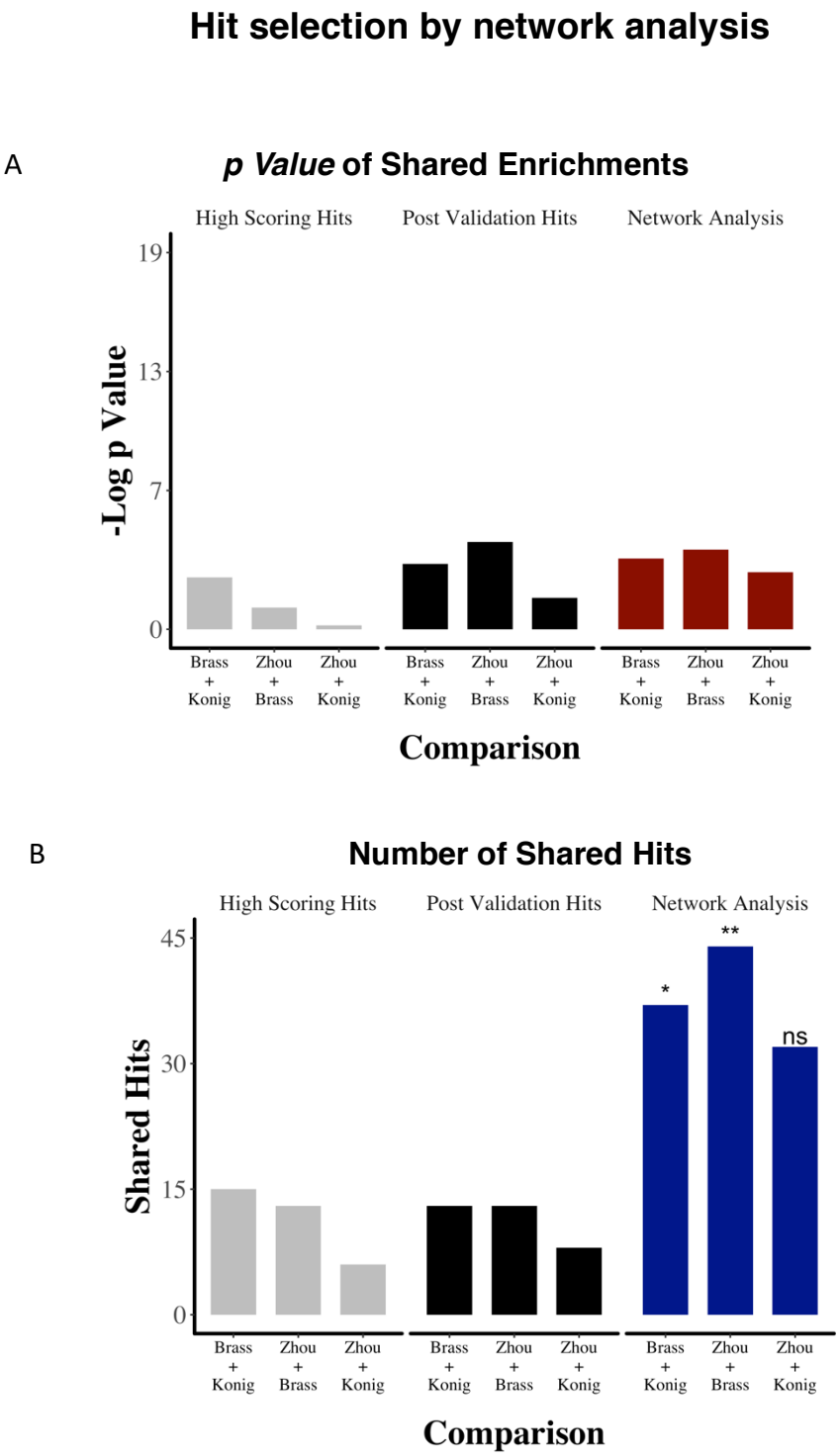


Figure 4.9: **Enrichment and overlap in hit selection by network analysis.**

(A) statistical significance of the overlap across the three studies of HDF when comparing hits selected by network analysis versus highest scoring hits and post validation hits. (B) Number of shared hits between the hits selected by network analysis from the three studies versus highest scoring hits and post validation hits. Random permutation test scores: ns = $p > 0.05$, * = $p \leq 0.05$, ** = $p \leq 0.01$

4.8 Pathway and network enrichment analysis are complimentary and non-overlapping in their solutions and hit selection methods

Results from the previous two sections suggest that pathway and network analysis provide alternative solutions to hit selection. Pathway analysis has the strongest impact on false positive correction (Figure 4.7A) and network analysis' impact is largely observed in false negative reduction (Figure 4.9B). These complementary solutions further suggest that an integrated framework that combines these two methods could be the optimal means to harness their combinatorial benefit to hit selection.

4.9 An integrated serial approach to pathway and network analysis improves both statistical enrichment and number of shared hits

To test a more integrated approach, I designed an integrated serial analysis framework for pathway and network analysis. I first identified enriched pathways from the high confidence hits of the three HDF screens, then promoted medium confidence hits that were members of the enriched pathways to high confidence. All high and medium confidence hits that were not part of the enriched pathways became the new medium confidence set. I then applied network analysis to the newly assigned high confidence and medium confidence sets, and again promoted any medium confidence hit that had an established interaction with a high confidence hit. The expanded set of high confidence hits were assigned as the final hit set (Figure 4.10).

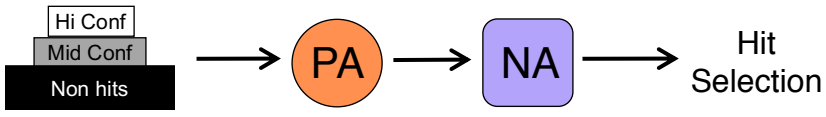


Figure 4.10: **Hit selection by serial analysis of pathway and network analysis.**

Schematic of the serial analysis approach for hit selection. Pathway analysis (PA) is applied to high confidence hits. High confidence and medium confidence hits from enriched pathways are assigned high confidence hits. Network analysis (NA) is applied to new set of high confidence hits. Medium confidence hits that have predicted interactions with high confidence hits are added to final hits.

Measuring the overlap of three screens of HDFs based on hit selection by the serial integrated approach led to improvements in both significance of overlap (figure 4.11A) and shared hits (figure 4.11B). While the improvements in significance and shared hit number were not as strong as with the respective exclusive use of pathway or network analysis, the improvement observed in both metrics when using the integrated serial framework suggests that it partially captures the combined error correction of the two methods.

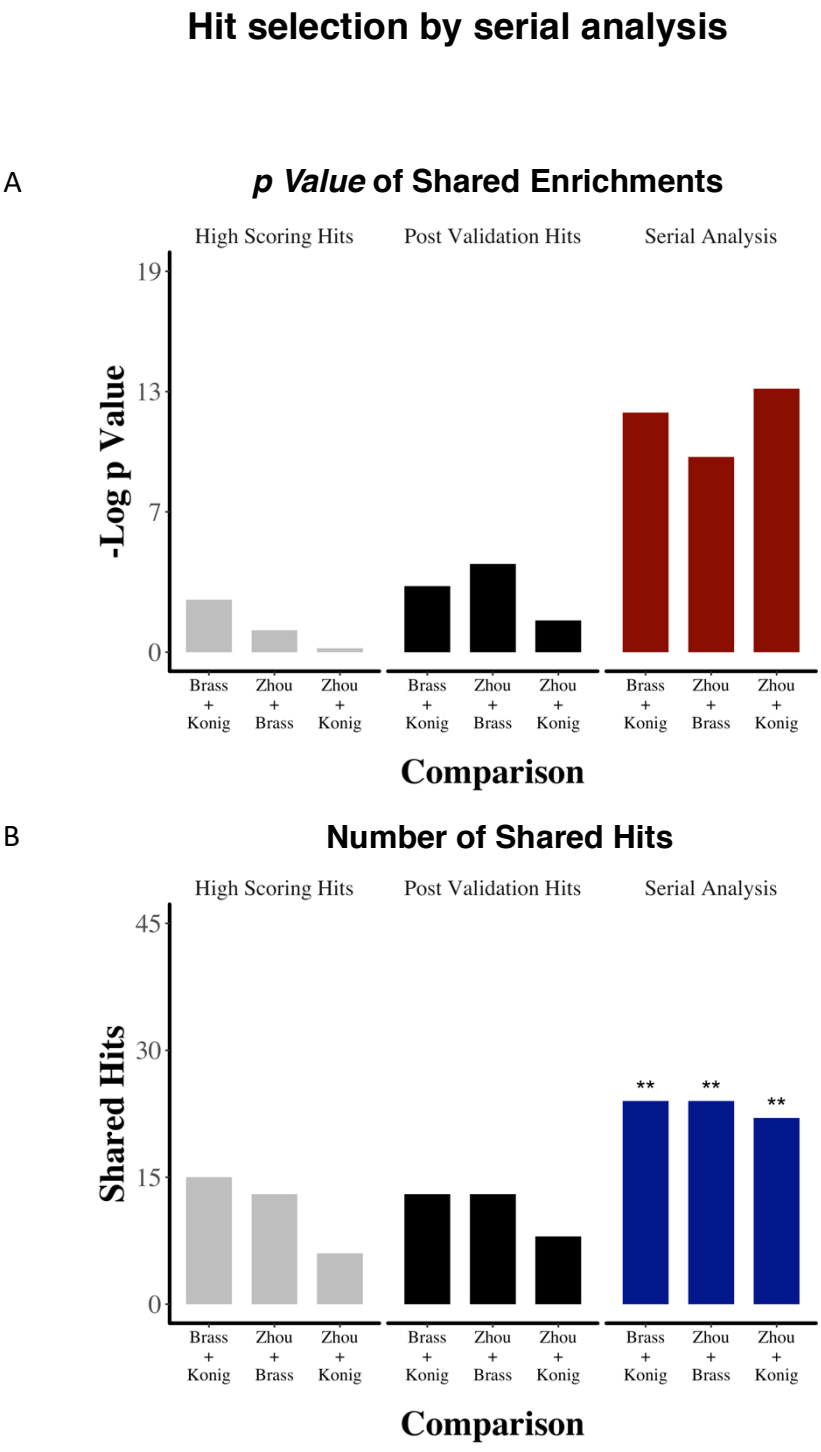


Figure 4.11: **Enrichment and overlap in hit selection by serial analysis of pathway and network analysis.**

(A) statistical significance of the overlap across the three studies of HDF when comparing hits selected by serial analysis versus highest scoring hits and post validation hits. (B) Number of shared hits between the hits selected by serial analysis from the three studies versus highest scoring hits and post validation hits. Random permutation test scores: ns = $p > 0.05$, * = $p \leq 0.05$, ** = $p \leq 0.01$

4.10 An iterative method for the integrated approach further improves on hit selection

In an attempt to amplify the combinatorial benefit of the integrated approach, I designed a framework that iteratively applies the two integrated analysis methods. In the iterative design, the same procedure as described above for the serial application of pathway and then network analysis is followed as a first iteration. Then, a second iteration repeats the pathway analysis step on the new set of high confidence hits, and the same procedure is followed to assign all high and medium confidence in the updated set of enriched pathways as high confidence hits. The same serial pathway-to-network analysis is applied to complete the second iteration.

If the set of high confidence hits has not changed in this second iteration the analysis terminates and the new set of high confidence hits is assigned as the selected hits. If, however, the new set of high confidence hits is not the same as the set of high confidence hits from the previous iteration, a new iteration of pathway-to-network analysis is applied. The iterative analysis therefore terminates when it identifies a high-confidence and medium confidence set of genes that are no longer changed by further analysis cycles (Figure 4.12). (In testing this method with a range of gene sets from different screens, I found that the analysis usually reaches an equilibrium after 5-6 cycles. The first iteration involving a contraction of the set of high confidence hits and the second iteration leading to an expansion of hits when pathway analysis is applied to the newly assigned high confidence set. After the second iteration only small number of hits are added with each iteration and after a few cycles small additions to the hit set no longer impact the enrichment, leading to an equilibrium in the hit selection of each step. See next paragraph and Figure 4.13 for an example of this). This approach ensures that the set of selected hits is modified until neither pathway or network analysis can pull it in a

different direction, ensuring that the resulting set of hits is at the equilibrium between false positive correction by pathway analysis and false negative correction through network analysis.

I then tested this iterative approach by applying the analysis to the three studies of HDFs. The three screens required 4 to 5 iterations before the set of high confidence hits no longer changed (Figure 4.13). I then ran the comparative analysis across the three screens using the hits selected by the iterative approach. Hit selection by the iterative integrated framework showed substantial improvements in both significance of overlap (Figure 4.13A) and size of shared hits (Figure 4.13B) across studies. Of note, the number of shared hits showed a marked improvement over the serial analysis method, suggesting that the repeated iterations are able to capture additional shared hits between screens that could be missed by a less rigorous analysis approach.

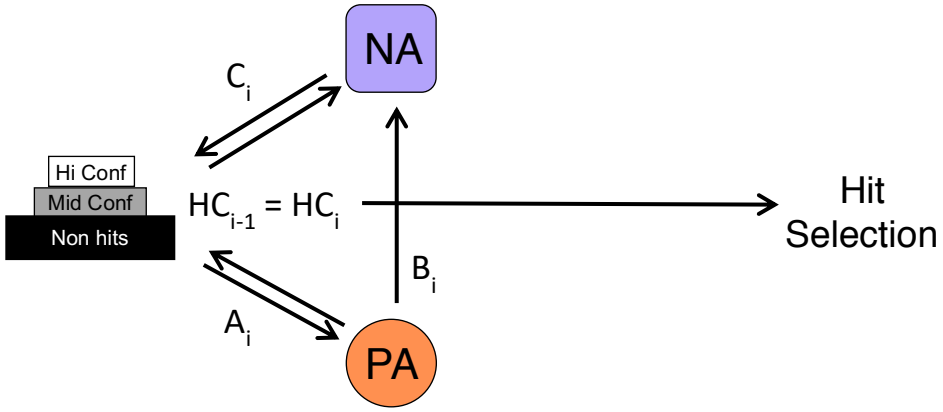


Figure 4.12: **Hit selection by iterative analysis of pathway and network analysis.**

Schematic of the iterative analysis approach for hit selection.
i = iteration 1
A_i: Pathway analysis (PA) is applied to high confidence hits.
B_i: High confidence and medium confidence hits from enriched pathways are assigned high confidence hits.
C_i: Network analysis (NA) is applied to new set of high confidence hits. Medium confidence hits that have predicted interactions with high confidence hits are assigned as high confidence hits.
Repeat steps A-C for i = iteration 2
When $i > 1$:
If the set of high confidence hits at the end of the current iteration (HC_i) is the same as the set of high confidence hits from the end of the previous iteration (HC_{i-1}) high confidence this are used as hit selection from the study. If high confidence set of of hits are different, repeat iteration.

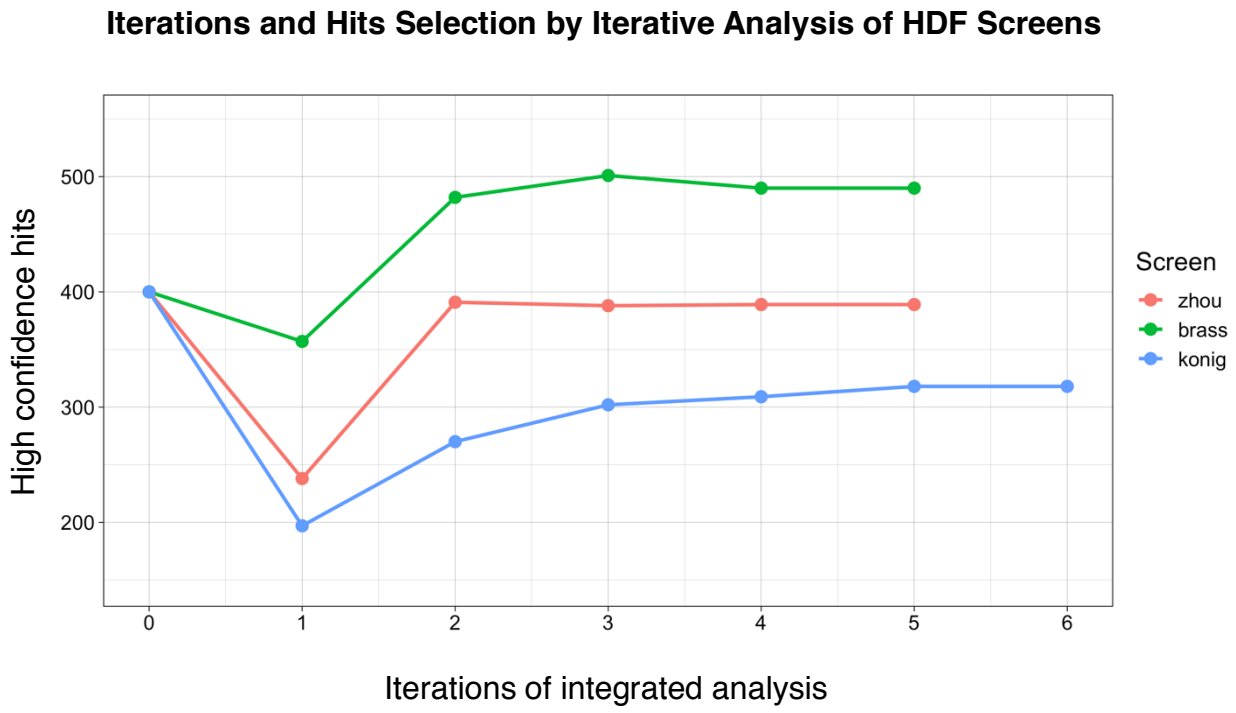


Figure 4.13: **Iterations of integrated analysis of the three studies of HDF.**

0 on the x-axis represents the high confidence set of hits at the input. The high confidence hit sets are contracted and expanded through iterative analysis. Analysis terminates when high confidence sets are similar between two consecutive iterations.

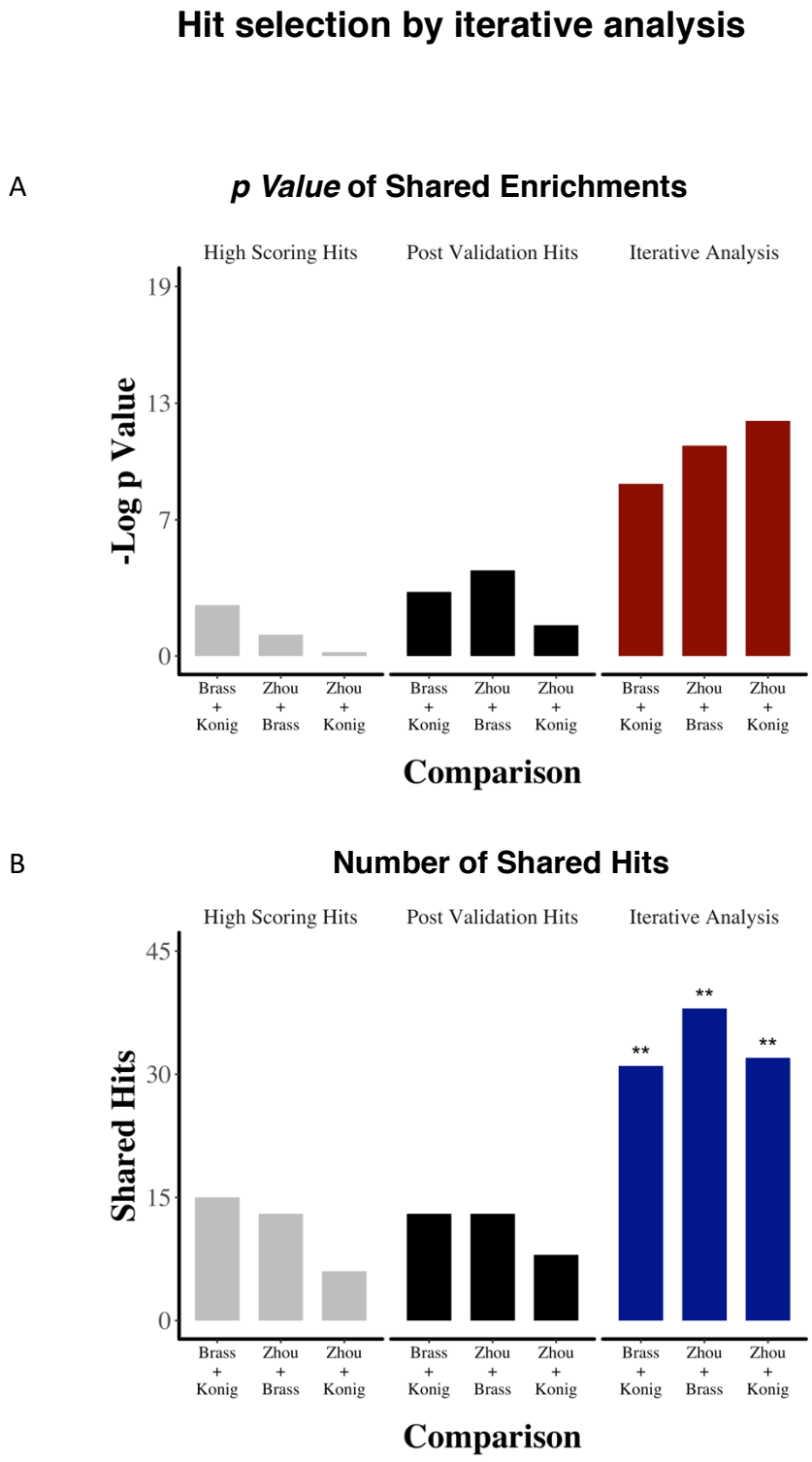


Figure 4.14: **Enrichment and overlap in hit selection by iterative analysis of pathway and network analysis.**

(A) statistical significance of the overlap across the three studies of HDF when comparing hits selected by iterative analysis versus highest scoring hits and post validation hits. (B) Number of shared hits between the hits selected by iterative analysis from the three studies versus highest scoring hits and post validation hits. Random permutation test scores: ns = $p > 0.05$, * = $p \leq 0.05$, ** = $p \leq 0.01$

4.11 Iterative pathway enrichment followed by network analysis outperforms alternative framework combinations

To ascertain whether the above design of the iterative framework is optimized to give the strongest improvements in hit selection, I designed and tested an iterative organization where the order of analysis methods is reversed, with network analysis done first followed by pathway analysis (Figure 4.15). The results of this analysis strongly aligned with the understanding of the complementary contributions of pathway and network analysis that I have outlined so far. While the same two analysis methods were applied, reversing the analysis order led to a decrease in both metrics of true positive hit selection (Figure 4.16A-B). The measure of confidence by the random permutation test was also low in two out of the three comparisons, suggesting that the noise of the false positive data was amplified in this hit selection approach. These results suggest that the preferred order for integrated analysis is a false positive correction (such as pathway analysis) followed by the false negative correction (as in network analysis), and that reversing the order greatly amplifies the noise of the results and blunts the power of the iterative approach.

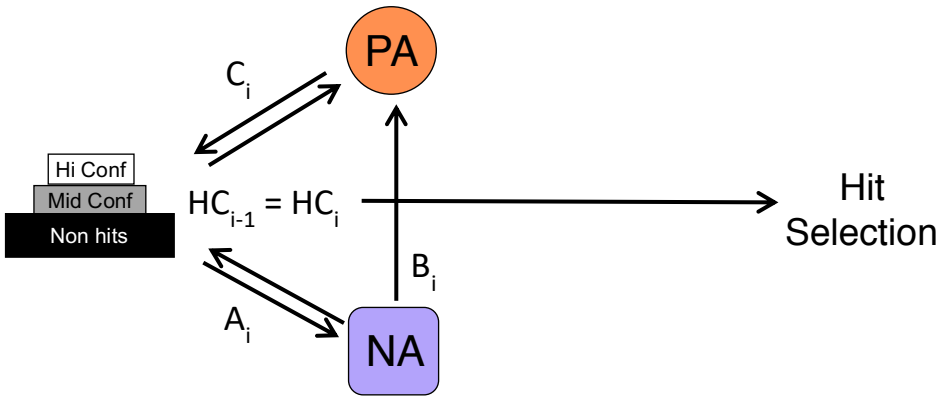


Figure 4.15: **Hit selection by iterative analysis with reverse pathway and network order.**

Schematic of the iterative analysis as in Figure 4.12 with the order of pathway and network analysis reversed.

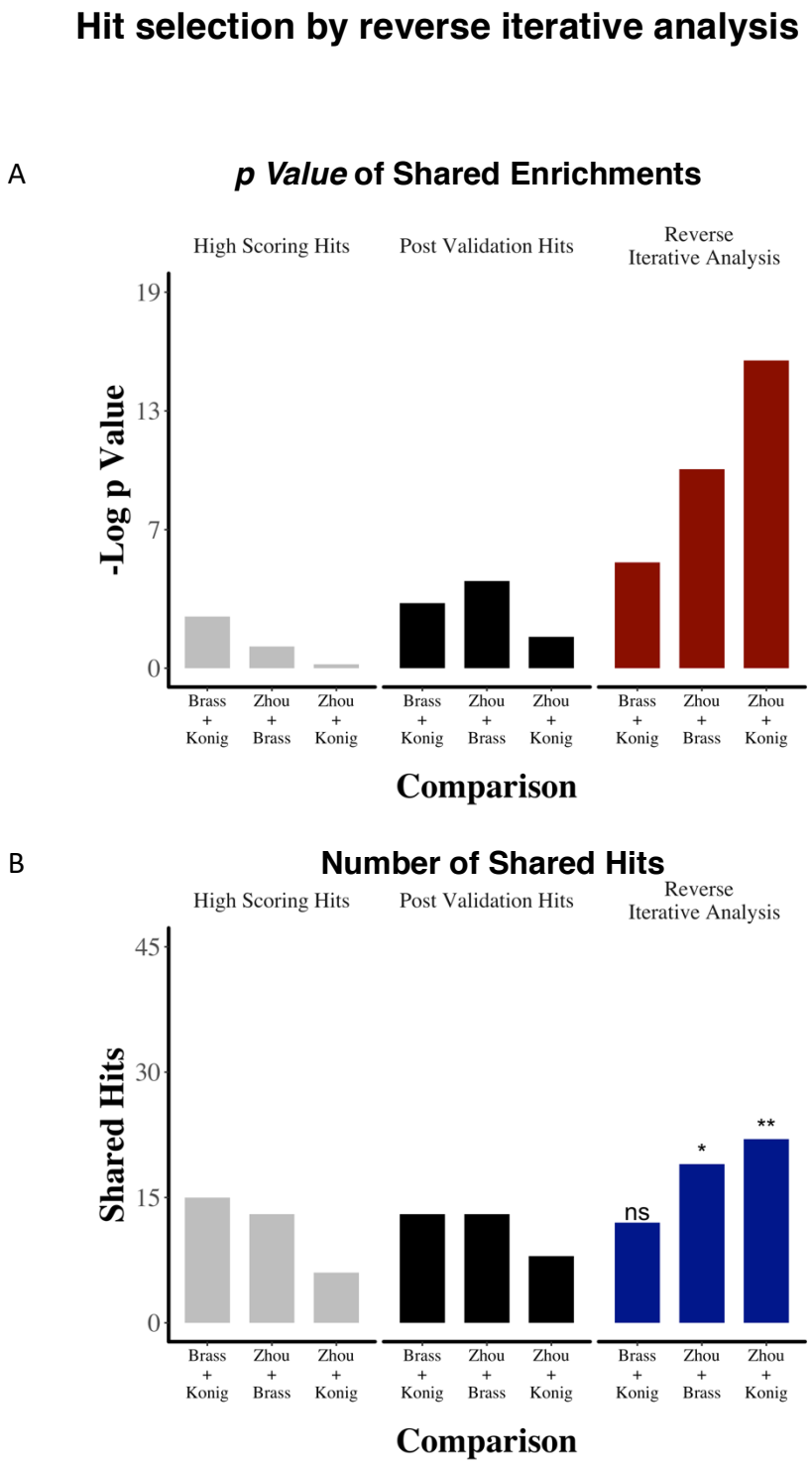


Figure 4.16: **Enrichment and overlap in hit selection by iterative analysis with reverse pathway and network order.**

(A) statistical significance of the overlap across the three studies of HDF when comparing hits selected by reverse iterative analysis versus highest scoring hits and post validation hits. (B) Number of shared hits between the hits selected by reverse iterative analysis from the three studies versus highest scoring hits and post validation hits. Random permutation test scores: ns = $p > 0.05$, * = $p \leq 0.05$, ** = $p \leq 0.01$

4.12 Throughput Ranking by Iterative Analysis of Genomic Enrichment (TRIAGE)

In a summary figure of the approaches tested and comparisons measured, I show that the iterative framework for pathway and network analysis provides the strongest combinatorial benefit of the complementary analysis approaches (Figure 4.17). By incorporating data that uses two cutoffs, this approach makes it possible to triage the results of a screen using a combination of the initial gene rankings from the screen and the known gene characteristics and functions from curated databases. Since this approach is not unlike the principle of medical triage as developed by the French physicians Dominique Jean Larrey and Pierre-François Percy in 1806 (Nakao et al., 2017), I chose the name TRIAGE for this approach as an acronym for Throughput Ranking by Iterative Analysis of Genomic Enrichment.

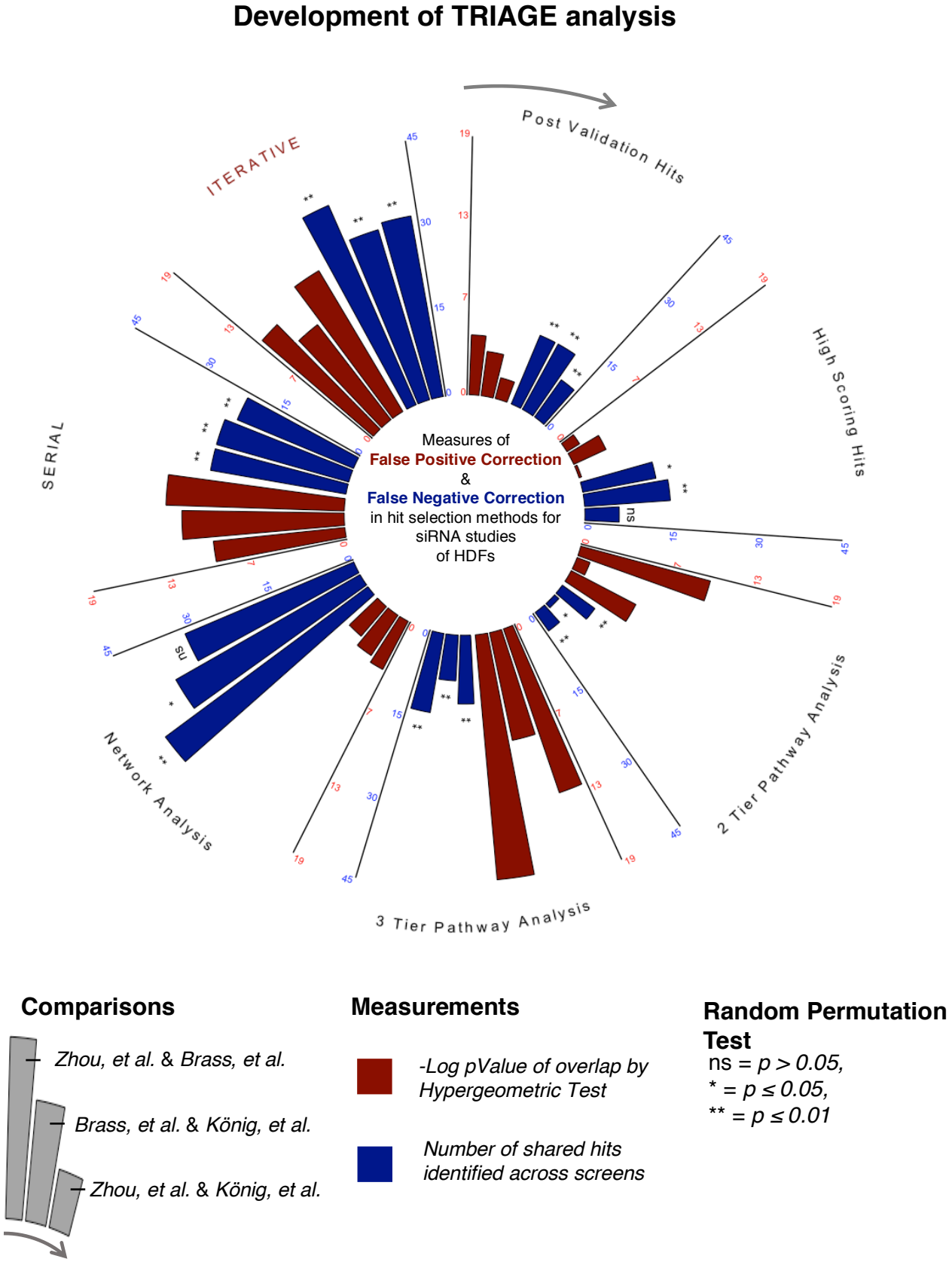


Figure 4.17: **Comparative analysis of different hit selection approaches.**

False positive correction measured by significance of overlap (red) and false negative correction measure by number of shared hits. Hit selection methods: (clockwise from top center): post validation hits, high scoring hits, pathway analysis using a two tiered dataset, pathway analysis using a three tiered dataset, network analysis, serial integration of pathway and network analysis, iterative integration of pathway and network analysis.

4.13 TRIAGE analysis can also be applied to post validation hits

I also tested TRIAGE analysis on the post validation hits from the three HDF studies to determine if the screen concordance was affected. I assigned as high confidence the reported hits from each screen. and medium confidence hits as the top 1000 genes not selected as hits by the study (Figure 4.18A-B). I then ran this tiered dataset through TRIAGE and observed similar improvements both in the number of shared hits across the screens (Figure 4.19A) and in the significance of overlap (Figure 4.19B). These results show that the TRIAGE hit selection approach can also be used for prioritizing hits from already published and analyzed screens. Through the appropriate assignation of high and low confidence hits, TRIAGE analysis can be applied to prioritize newer hits from older studies guided by the results of the previous analysis.

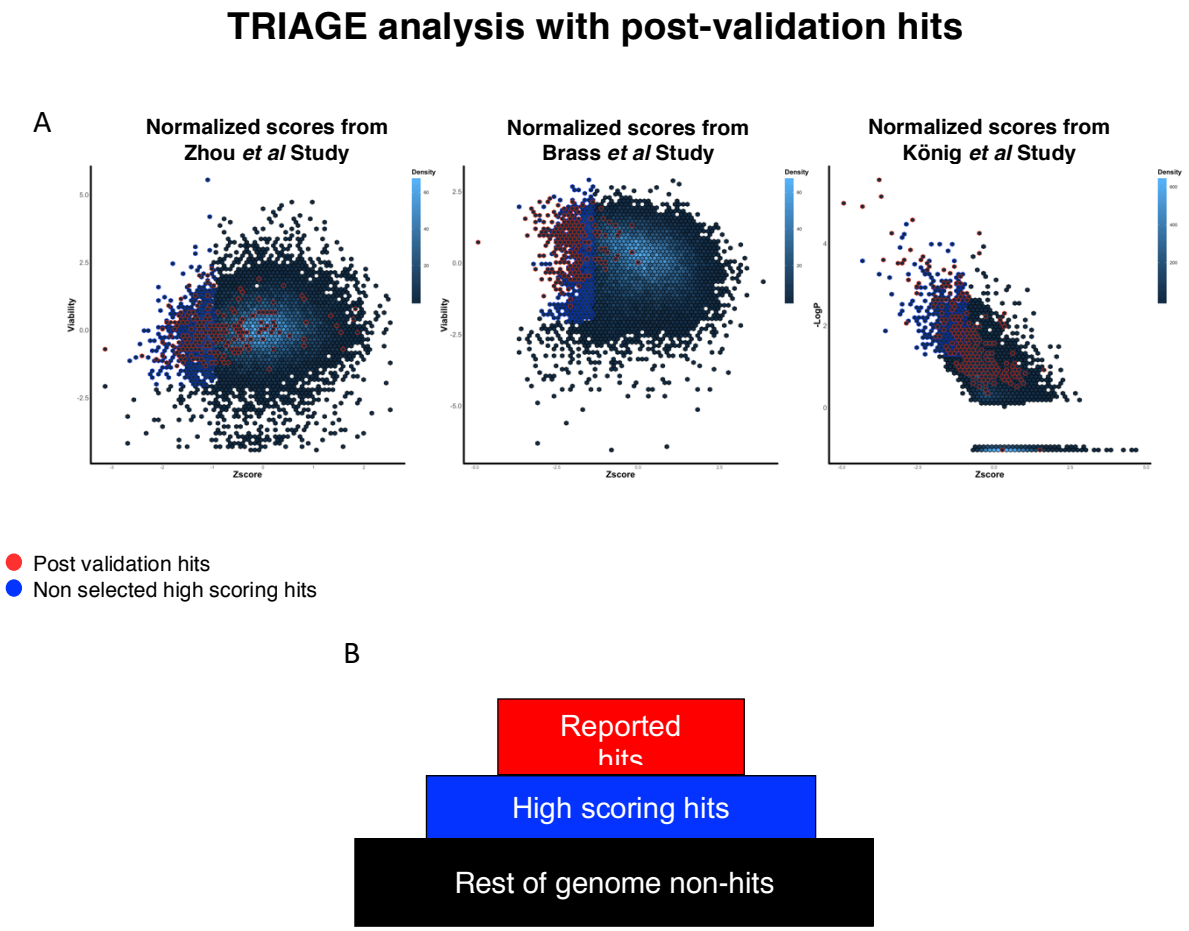


Figure 4.18: **Using post-validation hits for analysis by TRIAGE.**

(A) Scores from three genome-wide studies of HDF. Normalized scores are plotted on the x-axis and secondary scores that were considered (such as cell viability and assigned p-values) are on the y-axis. Genes reported as post validation hits are in red and 1000 genes with the highest Z scores that were not selected as post validation hits are in blue. (B) A schematic of the data tiering approaches using post validation hits. Reported post validation hits are assigned as high confidence. Highest scoring hits not selected by post validation are assigned as medium confidence.

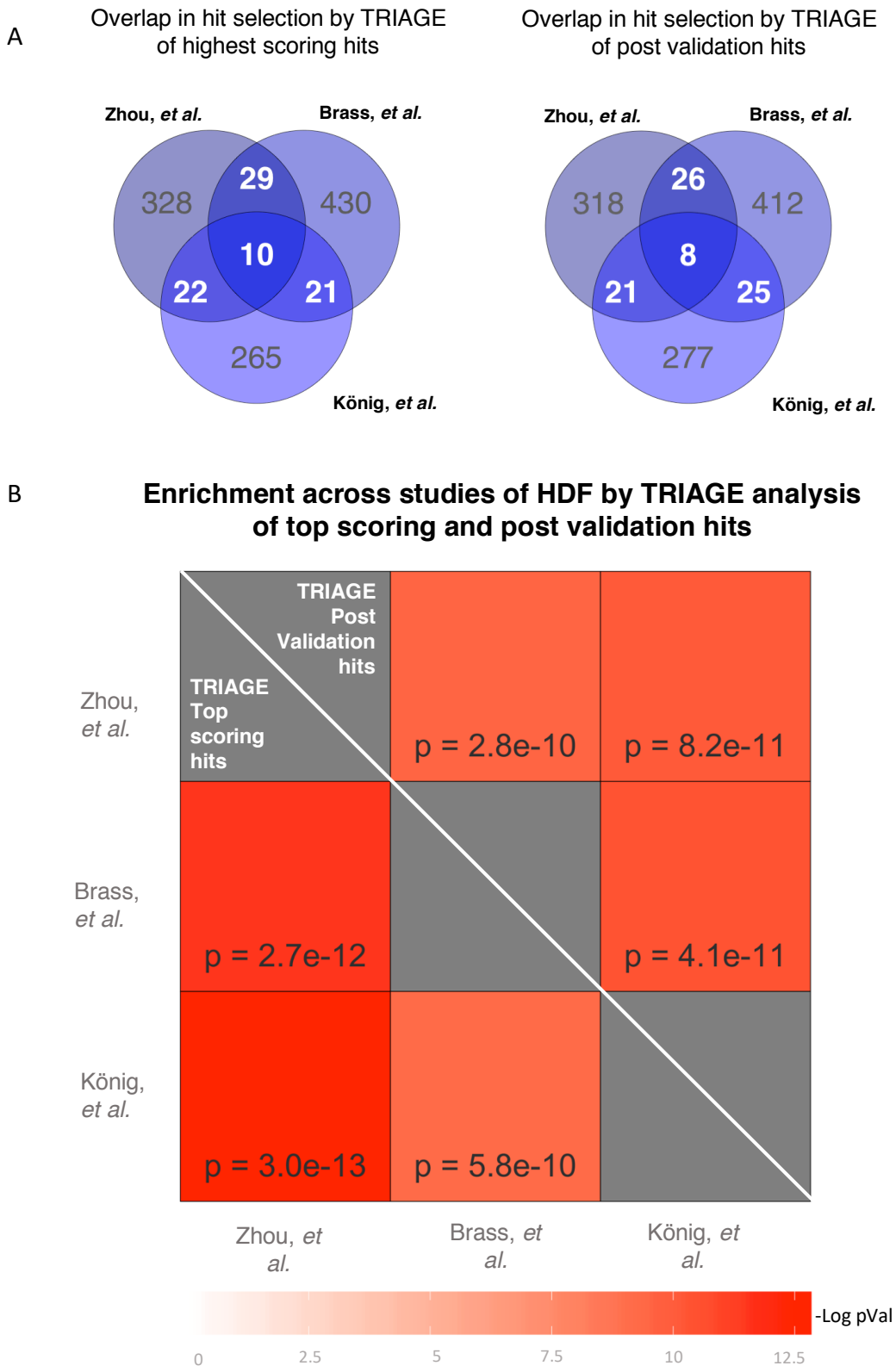


Figure 4.19: **TRIAGE analysis of highest scoring and post validation hits.**

(A) Venn diagram of the number of shared hits between hit selection sets by TRIAGE of highest scoring hits (left) and TRIAGE of post validation hits (right). (B) Analysis by hypergeometric test of shared enrichment across three studies of HDF by TRIAGE of highest scoring hits (bottom left) and by TRIAGE of post validation hits (top right). Shading represents –Log *p* value of the overlap.

4.14 Summary

In this chapter I have proposed a framework for how pathway analysis and network analysis can be integrated for improved hit selection from high-throughput studies. I demonstrated how this approach outperforms the application of pathway and network analysis individually or in a non-iterative serial design. This approach is not without its limitations, some intrinsic biases of database dependencies and enrichment biases persist and remain largely unavoidable with current database genome coverage and data content. An element of hit selection will also always rely on the judgment and work of the researchers. This design expands the possibility, however, of using analysis methods to capture more lower scoring hits and increasing the number of true positive results. In chapter 6 I will use this method to reassess the three studies of the macrophage response to LPS. In the design of the TRIAGE method, however, I was aware that this framework for hit selection can be beneficial to the screening community beyond those working on the TLR4 response. In the next chapter I will show how I made this approach adaptable to other datasets and available to a wider community of researchers.

5 A publicly available web-interface for TRIAGE analysis of high-throughput data *(triage.niaid.nih.gov)*

5.1 Introduction

As shown in chapter 4, the Throughput Ranking by Iterative Analysis of Genomic Enrichment (TRIAGE) method integrates analysis from network databases of known interactions and associations with the shared membership of genes in functional pathways to correct for the false negative and false positive error rates that are associated with score based hit selection methods. While I ultimately developed TRIAGE for analysis of the genome-wide studies of the LPS response in macrophages, the approach is also applicable to prioritizing gene selection from essentially any genome-scale study. In this chapter I discuss how I built a publicly available web-based platform that enables other researchers to quickly and easily prioritize candidates from omics studies using the TRIAGE approach (sections 5.2-5.5). The platform can be accessed globally (https://triage.niaid.nih.gov) and requires no prior knowledge of computational languages. I focused on making the platform adaptable to different types of datasets and results (sections 5.6-5.10). In this chapter I describe the steps of uploading a dataset to TRIAGE and how the features of the platform can be used to explore the data (sections 5.11-5.14). I also propose an alternative approach for how to visually represent network and pathway results together and build into the platform an interactive feature that helps guide exploration into 'missing links' between enriched pathways (sections 5.15-5.16). I also discuss the steps I took to ensure security and privacy of the data uploaded to the platform (section 5.17).

In developing the TRIAGE web interface, I collaborated with Dr. Jian Song from the Laboratory of Immune System Biology at the National Institute of Allergy and Infectious Disease (NIAID) of the National Institutes of Health (NIH) in Bethesda, Maryland, USA. In building the interactive network features on TRIAGE, I received assistance from Kyle Webb,

an NIAID Data Science Fellow. I also worked with the NIAID Office of Cyber Infrastructure and Computational Biology (OCICB) to secure the hosting and security measures for the site.

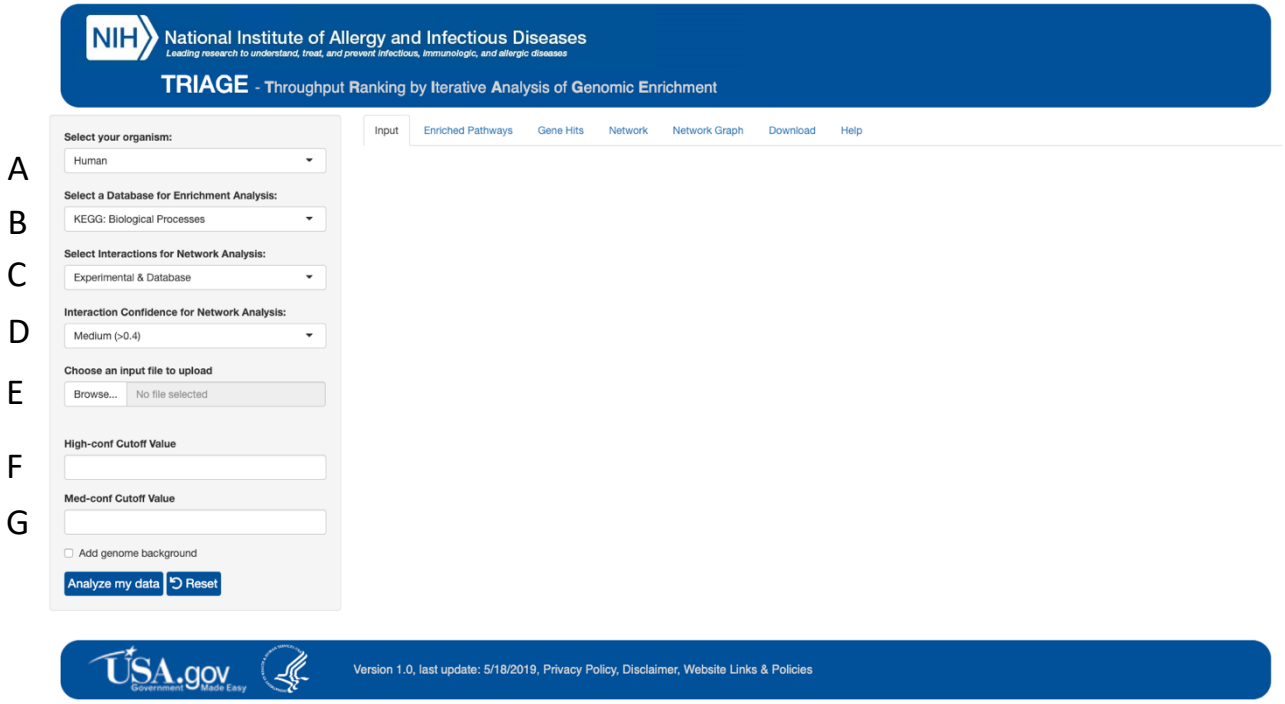
5.2 A Shiny driven web interface for TRIAGE analysis

I developed and designed the TRIAGE analysis pipeline using the "R" computer language. R is a computer language and environment designed for statistical analysis and data visualization (R Core Team, 2014) and is one of the top three most commonly used platforms in computational biology (together with the Unix shell Bash and Python) (Carey & Papin, 2018). Data analysis solutions created with R can be reproduced by either shared code or through the sharing of "packages", units of self-contained reproducible code that can be integrated into the workflow of other environments (Wickham, 2015). Despite these advantages, however, there remains a critical need within the omics community to make analysis pipelines available to researchers across the spectrum of computational literacy. One such path is utilizing the R package "Shiny" which makes R based scripts run as interactive web interfaces that can be used without having to interact with or manipulate computational code (Winston Chang, 2019). Examples of R based analysis pipelines that have utilized the Shiny platform to create widely used web-based interfaces include the Comprehensive Analysis for RNAi Data (CARD) platform (B. Dutta et al., 2016) and the Pathway Coexpression Network (PCxN) platform (Pita-Juárez et al., 2018).

To build a Shiny powered interface for TRIAGE and broaden the utility of the platform, I separated what choices can be provided to each user and what features are preset to apply to all analyses. A table of the preset conditions versus user provided inputs is listed (Table 5.1). Based on these settings I built a Shiny powered interface that can record the user provided inputs as parameters for running the analysis. (Figure 5.1)

	Pre-set	User-selected
Organism		Human or Mouse
Pathway Database	KEGG	Biological process pathways, disease pathways, or all pathways,
Network Database	STRING	Which ”evidence source” interactions to include. Which confidence level of interactions to include.
Criteria for high confidence, medium confidence and non-hits		Selected from a drop-down menu created from the input fields in the user’s upload file
Cutoffs		Numerically assigned by the user
Statistical threshold for enrichment	Within the TRIAGE analysis the cutoff is set at $p = 0.05$	Following analysis, user is provided with p values, FDR, Bonferonni score for all enrichments

Table 5.1: **Pre-set and user selected conditions in TRIAGE interface.**



A
B
C
D
E
F
G

Figure 5.1: ***triage.niaid.nih.gov*** **interface.**

The landing page of *triage.niaid.nih.gov.*
(A) Users selects the organism relevant to the dataset (human or mouse). (B) a drop down menu to select the relevant pathway types from the KEGG database. (C) The default setting for network analysis uses experimental and database sourced connection. The user can also select "advanced options" which creates a list of all the possible interaction types that the user can select from. (D) The user can select the confidence cutoff for network interactions based on the scores from STRING. Default is set to medium confidence. (E) .csv files can be uploaded to the platform by clicking the browse button and selecting the file. (F) Field for entering the cutoff or assigned value for high confidence hits. (G) Field for entering the cutoff or assigned value for medium confidence hits.

5.3 Unique IDs can be assigned by either EntrezID or HGNC Gene Symbol

Commonly used gene names in the scientific literature use the familiar "Gene Symbol" nomenclature. While a gene symbol is often more familiar and intuitive to interpret, the NCBI EntrezID numeric nomenclature is a more reliably unique identifier with a one-to-one matching of gene to ID. The TRIAGE analysis uses EntrezID to ensure consistent and accurate gene matching between databases and user input files. The front facing information in TRIAGE, however, uses GeneSymbol to make the information more interpretable by the user. To that end, the files uploaded to TRIAGE for analysis must include a column assigned either "EntrezID" or "GeneSymbol", with the appropriate ID matched to each input (Figure 5.2). Both ID systems, however, do not need to be included in the upload. If only one is included, I designed the platform to identify the missing one and generate a column of the missing ID type.

5.4 KEGG pathway database for human and mouse is integrated and adapted in the TRIAGE platform

To adapt the KEGG database to the TRIAGE platform I included the pathway membership data from the KEGG Application Program Interface (API). I filtered the pathways into two groups "Biological Processes" and "Disease Pathways". In the TRIAGE interface, the user can select whether to use one or both of these groups when querying the data for enrichment (Figure 5.1B)

GeneSymbol	EntrezID	PercInfected.Zscore	CellNumber.Zscore	assigned.value
CXCR4	7852	-4.96623446	0.755277194	1
C1orf52	148423	-3.637572822	1.925896515	1
MED14	9282	-3.435696475	0.976086257	1
ADAM10	102	-3.435673997	1.513583032	1
GCK	2645	-3.223640638	1.920937941	1
GPR21	2844	-3.201246647	-0.096137148	1
ZNF831	128611	-3.20098761	-0.243515999	1
CD4	920	-3.162525347	1.767687649	1
EGFR	1956	-3.162525347	1.487870985	1
WNT1	7471	-3.140163554	1.674727332	1
USP6	9098	-3.114415882	1.238823011	1
PLEKHA7	144100	-1.920126088	-1.705268716	0.5
DPH3	285381	-1.919642212	-0.473272261	0.5
NA	284861	-1.919642212	-0.200375892	0.5
PNMA6A	84968	-1.917641072	-1.574602053	0.5
EIF3G	8666	-1.917428352	-1.51997327	0.5
TFDP2	7029	-1.916264377	1.3078323	0.5
CLNS1A	1207	-1.915660931	0.543313511	0.5
MMP19	4327	-1.90908986	1.426094148	0.5
RECQL4	9401	-1.90908986	1.167972495	0.5
ZNF536	9745	-1.909010831	-0.111569716	0.5
NMUR2	56923	-3.690591512	-2.63385185	0
SMU1	55234	-3.686455305	-3.076050904	0
LSM8	51691	-3.62462392	-2.307921727	0
NAT10	55226	-3.460889184	-2.681841646	0
SGO1	151648	-3.435116303	-4.066983289	0
DHRS13	147015	-3.38711184	-2.495038806	0
XAB2	56949	-3.327710602	-3.064859872	0
HEG1	57493	-3.291433863	-2.728928693	0
COPB2	9276	-3.28475593	-4.284378793	0
PSMB6	5694	-3.261635121	-3.280551426	0

Figure 5.2: **A sample input file for TRIAGE.**

A sample dataset prepared for TRIAGE analyses using the data from the *Brass et al.* of study of essential factors for HIV infection. Gene column IDs are labeled as "EntrezID" and "GeneSymbol" (either one is sufficient for upload). The "PercInfected.Zscore" column includes the normalized Z scores and can be used to set cutoffs for in the high confidence and medium confidence fields on the TRIAGE platform. To include the "CellNumber.Zscore" to consider what is a high confidence and what is a medium confidence hit, a new column is created "assigned.value". Hits assigned as high confidence by both criteria are given a value of 1, hits assigned as medium confidence are given a value of 0.5. Hits that don't meet the two criteria are assigned a value of 0.

5.5 STRING database interaction networks and criteria are integrated into the TRIAGE platform

Adapting the STRING database for use by TRIAGE required first an alignment of the IDs, as STRING uses Ensemble protein IDs and TRIAGE uses gene EntrezIDs. Using the Biomart converter table, I transformed the STRING network data by mapping the proteins IDs to EntrezIDs (Durinck et al., 2009). I then segmented all interactions based on their evidence source and divided the confidence cutoffs into three tiers; 0.15-0.4 as low confidence, 0.4-0.7 as medium confidence, and 0.7-1 as high confidence. This allows the user to choose and compare different networks and confidence levels for their analysis. As the preferred settings for high-throughput hit selection are interactions in STRING that are based on experimental and database evidence of at least medium confidence, I set these as the default TRIAGE settings. I also included as an advanced option the ability for users to select their preferred evidence sources and confidence level (Figure 5.1C-D). The filtered network set by the user at the outset of the analysis is used throughout all of the subsequent analysis steps.

5.6 TRIAGE platform can perform analysis with human or mouse datasets

To broaden the utility of the TRIAGE platform I designed the analysis with a built-in capability to run analysis for mouse or human datasets. When setting the analysis, the user selects the appropriate organism in which the data was generated, and all the subsequent database selections are matched accordingly (Figure 5.1A).

5.7 Uploading a data set to the TRIAGE platform for analysis

The TRIAGE platform reads files in the .csv format. The file must contain a column either titled “GeneSymbol” that has HGNC gene symbols in all the rows or, alternatively, a column titled “EntrezID” with NCBI EntrezID in all the rows. The file must also contain a column with a numeric value that can be used to separate high confidence, medium confidence, and non-hits. This column can be titled according to the user’s choosing. A sample dataset is shown in Figure 5.2. After the file is uploaded a message appears with the number of IDs in the list that have been successfully mapped to gene IDs with a warning of the number of IDs that are unmatched

5.8 Criteria for high confidence and medium confidence cutoffs are set by the upload file and user

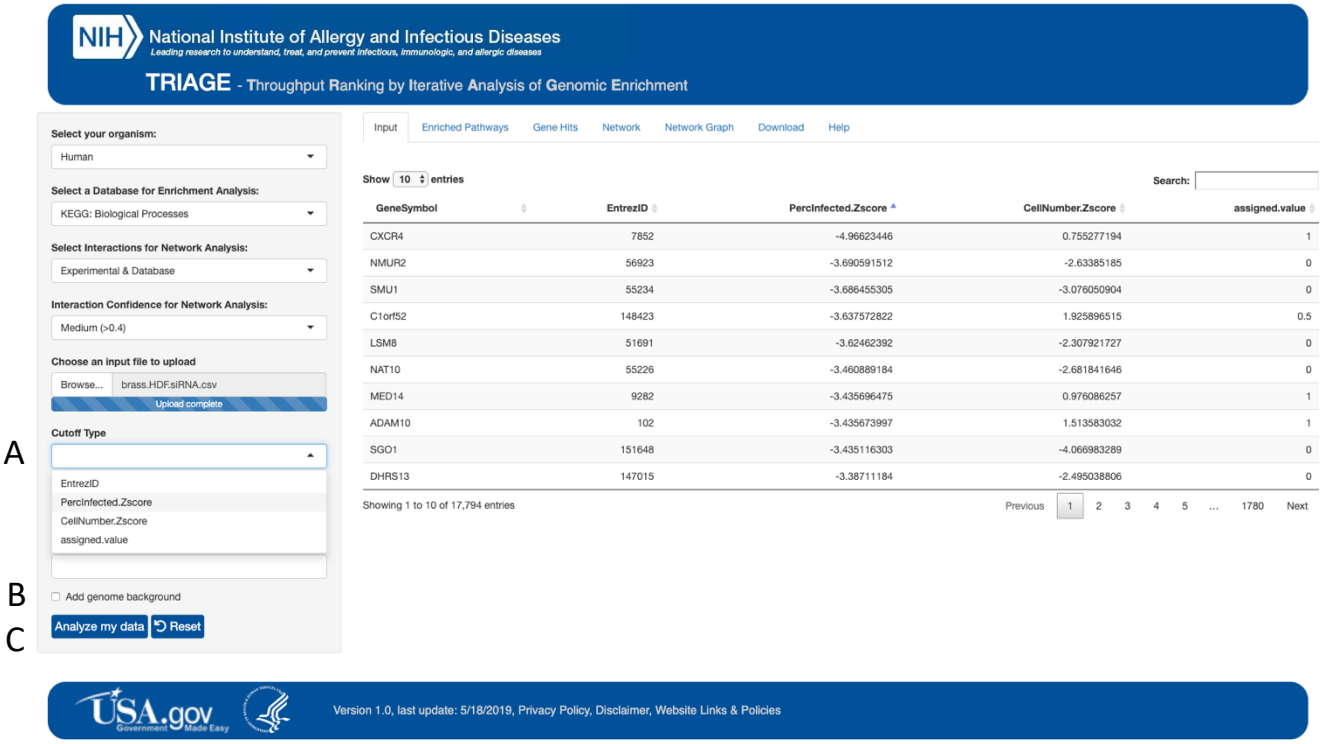
After the user’s file has been uploaded a dropdown option appears under ‘Cutoff Type’ containing the names of all the columns in the upload file (Figure 5.3A). The user can then select the name of the column that contains the numeric values to be used for setting the high confidence/medium confidence cutoffs. For the rest of the analysis only the columns containing the gene IDs and the selected “Cutoff Type” column will be considered by the platform, any additional columns in the upload file are ignored and added back only in the download file that the user saves at the termination of the analysis.

5.9 Dual cutoffs in TRIAGE platform can be set by assigned values or as greater-than or less-than values

For simplicity, I designed TRIAGE to consider the input of only one column when measuring what is a high confidence hit and what is a medium confidence hit. In a straightforward way

this column can be used for a measure like Z score, fold change, or *p* value. The user then places the numeric cutoffs for high confidence and medium confidence hits in the appropriate field of the interface. TRIAGE then automatically calculates, based on the difference between the cutoffs for high confidence and medium confidence, whether it should be using a greater-than or less-than approach when segmenting the data by cutoff.

This design, however, does not preclude using the inputs from multiple criteria to assign targets to high confidence or medium confidence groups. When determining what should be designated as high confidence and medium confidence, a user might want to consider a primary readout (such as Z score or fold change) together with a secondary readout (such as including cell viability measures, *p* values, off target analysis, etc.). Some user may also want to combine datasets from different studies to create confidence designations based on the scores form different assays. To consider more than one criterion a user can do the confidence assignation before uploading the dataset to TRIAGE (such as seeing which hits meet more than one criterion, which hits appear in more than one assay, etc.) and assign all of the high confidence hits a single value (such as 1 for example). The user then assigns a value that is below the value chosen for high confidence hits (such as 0.5) to all the hits that would be considered medium confidence. The user then assigns a value below both of these values to all other hits (such as 0). The user then uploads the file and enters "1" as the high-confidence cutoff and "0.5" as the medium confidence cutoff (Figure 5.1F-G). This will group all the hits curated from the different criteria in the proper confidence designations. Examples of these different approaches are shown in the sample dataset in Figure 5.2.



A

B

C

Figure 5.3: **Running TRIAGE analysis.**

(A) After a file is uploaded a dropdown menu with all the column names containing numeric values appears. The user selects the preferred column for setting cutoffs. (B) When selected, "Add genome background" adds in a 'rest of genome' background of non-hits for analysis of the dataset. (C) Analysis can be run and reset within the same session.

5.10 TRIAGE can add a "genome background" for datasets that only include hits

Many high-throughput assays are now outsourced to core facilities within institutions and then sent to research groups to analyze and follow up on. These resources often supply in return a list of selected candidates and do not include the full range of scores for the entire genome being measured. These restricted lists preclude the possibility of running enrichment statistics that measure the number of hits against the entire set from which they were selected from. To not exclude the analysis of those list's formats on TRIAGE, I added a feature where the user can add in a "genome background" against which the enrichment analysis will be done (Figure 5.3B). When selected, the "add genome background" feature adds genes to the list that aren't included in the upload file to be used as a background for statistical enrichment analysis. The added background "genes" will not appear as suggested hits by the TRIAGE analysis, the background genes are only used as a means to have more robust statistics on the enrichment of pathways. The background genomes use only the known protein coding genes of the selected organisms that are not in the upload file. TRIAGE uses the difference between the size of selected hits and the number of known protein coding genes as the number of "non-hits" for enrichment statistics.

5.11 Running TRIAGE analysis with a single click and in less than 30 seconds

To run TRIAGE analysis, the user clicks "Analyze my data" and the iterative analysis process begins (Figure 5.3C). A progress bar appears once the analysis has begun and makes half step leaps after every iteration so that the user can see the progress. In my testing, a majority of datasets complete the analysis in less than 30 seconds. When the analysis has run, the window switches to the results panels (see following sections). An added benefit of the fast speed of the analysis processes is that it allows a user to experiment with different cutoffs for the high

and medium confidence settings to compare and contrast outputs. To encourage this approach, I included a “Reset” icon that, when clicked, resets the input settings and allows the user to easily run a new analysis with different parameters (Figure 5.3C).

5.12 TRIAGE has built-in recognition for contingencies and outlier datasets

It is impossible to predict all the different ways a web-based analysis tool will be used. Not all of the directions provided will be followed and there are a myriad of ways in which input files may vary from the ones for which I originally built this platform. For situations that may arise when some of the input and upload criteria are not fully followed, I took the approach of building a warning message system that alerts the user as to how an input setting led to an incomplete result. This approach has the benefit of guiding the user through their attempts to correct their input settings. Table 5.2 lists the anticipated situations and the associated warning messages I included.

While testing multiple randomly generated datasets on TRIAGE, I identified rare cases where the enrichment set is so small, that the statistics for enrichments are extremely sensitive to small additions to the sets of hits. These situations lead to the set never fully converging to the same set of hits but instead fluctuating between two sets of hits. To anticipate the possible emergence of these cases I added in a contingency whereby the iterations of TRIAGE first check to see if the set has converged and if it hasn’t it looks for a repeated pattern. If the iterations are found to oscillate between the same repeated selections of sets, it looks for the largest most inclusive set in the pattern and terminates the analysis at that juncture.

Anticipated Error	Warning Message
Wrong organism selected or wrong Gene Symbol convention used	Either check the organism or update your GeneSymbols to match the official HGNC symbols if you want to include ALL in this analysis.
One or more input genes have no matching EntrezID due obsolete GeneSymbol	Warning: X GeneSymbols have mapped EntrezIDs and will be used in this analysis!
Criteria generate too small datasets for network analysis	Warning: Criteria produced empty network. Session will restart.

Table 5.2: **Anticipated user errors with built in responses.**

5.13 TRIAGE analysis identifies robust enrichment and illustrates annotated KEGG pathway maps

Once an analysis has run the window switches to the "Enriched Pathways" tab. This tab provides a list of statistically enriched pathways found in the selected set of hits by TRIAGE analysis. The list includes all pathways that have a statistical score of 0.05 or less in a hypergeometric test. While a 0.05 *p* value is a lenient cutoff for pathway enrichment, this is only the minimum requirement a pathway must meet to be included in the list. Values for more stringent statistical tests are included in the table and allow the user to rank pathways based on different standards and cutoffs (Figure 5.4).

The names of the enriched pathways can also be clicked on. Clicking on the pathway name will open a new tab from the KEGG website showing a schematic of the genes in the pathways with the gene hits from the TRIAGE analysis highlighted. Genes that were marked as high confidence at the start of the analysis are highlighted in blue and those marked as medium confidence are highlighted in red. This feature makes it possible to further explore if the genes that are driving the enrichment of the pathway are spread across the pathway or concentrated in a particular segment (Figure 5.5)

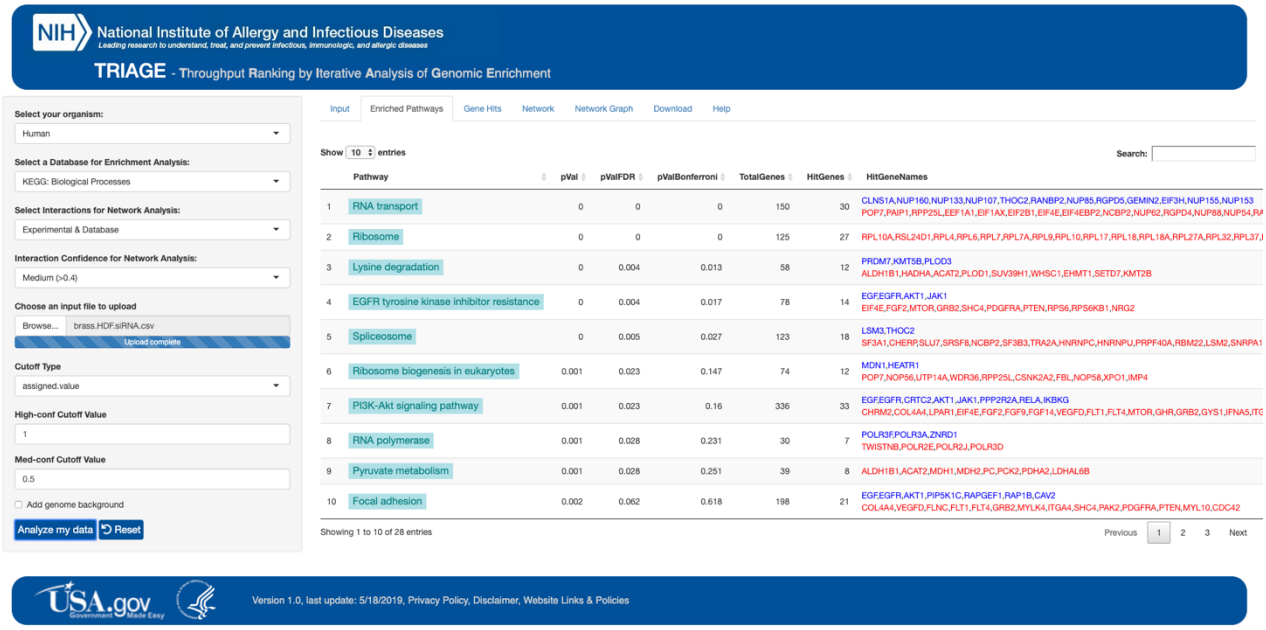


Figure 5.4: ***triage.niaid.nih.gov*** **results.**

A table of enriched pathways following TRIAGE analysis of the *Brass et al.* study of essential factors for HIV infection. Hit genes selected by TRIAGE that were in the input high confidence group are in blue, hit genes selected by TRIAGE that were in the input medium confidence group are in red.

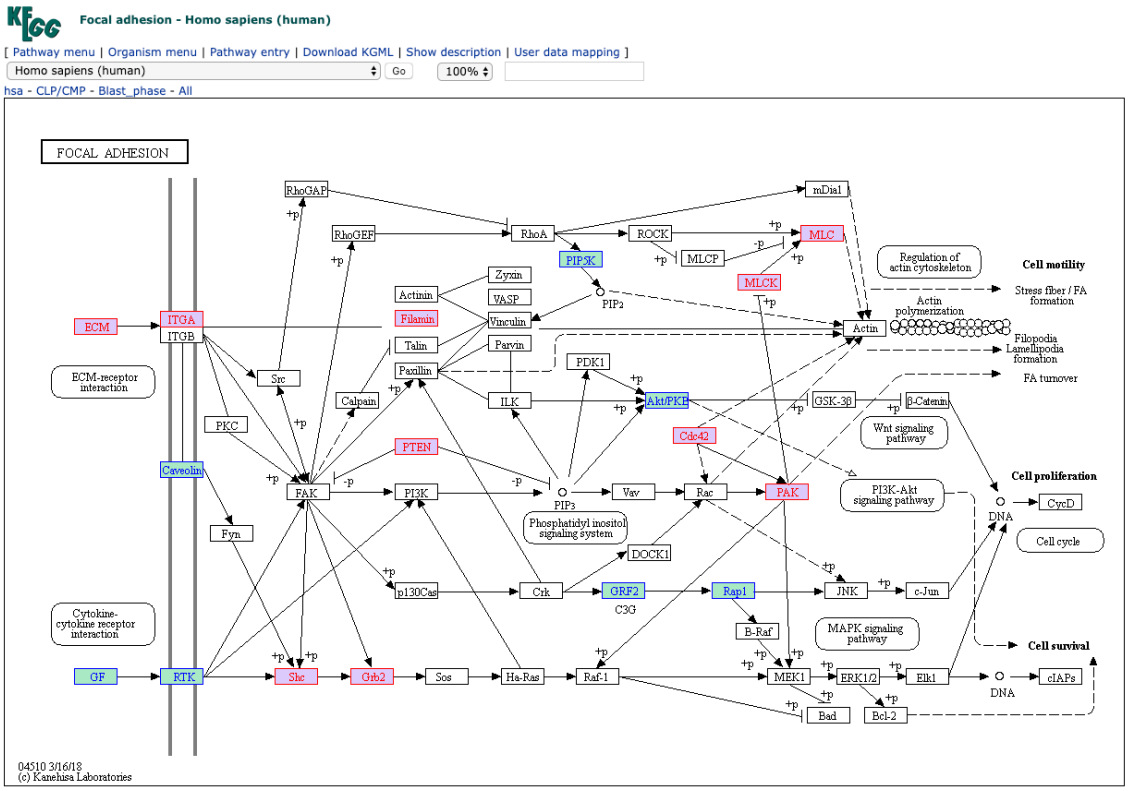


Figure 5.5: **Results from TRIAGE analysis mapped onto a KEGG pathway map.**

Focal adhesion pathway map overlaid with the results from TRIAGE analysis of the *Brass et al.* of study of essential factors for HIV infection. Hit genes selected by TRIAGE that were in the input high confidence group are in blue, hit genes selected by TRIAGE that were in the input medium confidence group are in red.

5.14 TRIAGE output provides prioritized hits from high-confidence and medium confidence sets

Following the interactive enrichment tab, the "Gene Hits" tab contains a series of tables that list the genes that were selected by the TRIAGE process and can guide the further prioritization of hits for follow up. The lists are divided under different sub tabs:

TRIAGE Gene Hits: This table provides a list of prioritized hits selected by TRIAGE analysis with supporting information on interacting genes (based on the user -selected network criteria) and membership in enriched pathways. I designed these tables to best guide further analysis. The "interacting gene" column lists which other hits each selected hit has predicted interactions with. An additional column lists the pathways that those interacting genes are members of. This can be particularly useful when the specific gene hit is not yet annotated as part of a pathway. Using the pathway memberships of its interacting genes can help guide hypotheses about its function.

Gene Hits by Iteration: A table with the input document and the genes added or dropped out at each iteration is listed in each column. Genes counted assigned as high confidence in a specific iteration are indicated as "HighConf", genes counted as medium confidence are indicated as "MedConf". The "TRIAGEhit" column indicates the gene hits that are counted as final hits in TRIAGE. The top row highlighted in orange indicates the total number of hits considered high confidence at the end of each iteration. The table also includes columns with information about the pathways, interactions with other hits, and the pathway membership of the interacting genes as in the TRIAGE Gene Hits table described above.

Graph: Gene Hits by Iteration: A graph showing the number of medium confidence hits and high confidence hits that are selected as TRIAGE hits at each iteration of the TRIAGE analysis. (starting with all the initial high confidence genes from the input and 0 medium confidence genes. Medium confidence hits are added during the iterations).

High Confidence Hits not in TRIAGE Hits: This table includes a list of hits that were assigned as high confidence in the input but were not selected as hits by the TRIAGE analysis. I added this table so that user can easily review the high-confidence hits that were dropped out by TRIAGE to see if any of those should be added back in. This is an important feature in the context of under studied genes that have no previously reported pathway annotations or interactions, but could be an important novel regulator of the biological process being studied.

Pathway Enrichments: This table lists the enriched pathways from the analysis with statistical cutoffs and the gene candidates that drive the enrichment.

The enrichment and hit prioritization tables are all available to download by the user. This ensures that the data is securely saved by the user after the analysis is removed from the server. The "Download all files" icon under the "Download" tab provides a zipped folder of the analysis files generated by the TRIAGE platform.

5.15 Using Hierarchical Edge Bundling (HEB) to simultaneously visualize pathway and network enrichment

In building the TRIAGE platform I encountered the challenge of how to visually represent the results from the combined pathway and network analysis. To build a solution I adapted a network visualization approach called Hierarchical Edge Bundling (Holten, 2006).

Hierarchical Edge Bundling is based on the design principle of all the nodes in the network being organized in a circle and separated into different groups (the hierarchical part), the edges connecting different nodes to one another across groups are then "bundled" such that you can observe both the individual node and edge origin. This visualization also enables you to see which groups are often connected (the edge bundling part). For this context I utilized the Hierarchical Edge Bundling method to group network nodes (which in this context would be genes or proteins) that are part of enriched pathways into individual groups. I assigned another group as the "novel gene" group to place all the gene hits that are not annotated as part of the selected pathways. For visual clarity, I filtered out the often seen intra-group connections within a given pathway, but kept them for the group representing the "novel genes". This method allowed for easier visualization of genes driving suggested interactions between pathways. This approach also makes it possible to explore interactions between genes in the same pathway through a common interacting 'novel' gene (Figure 5.6).

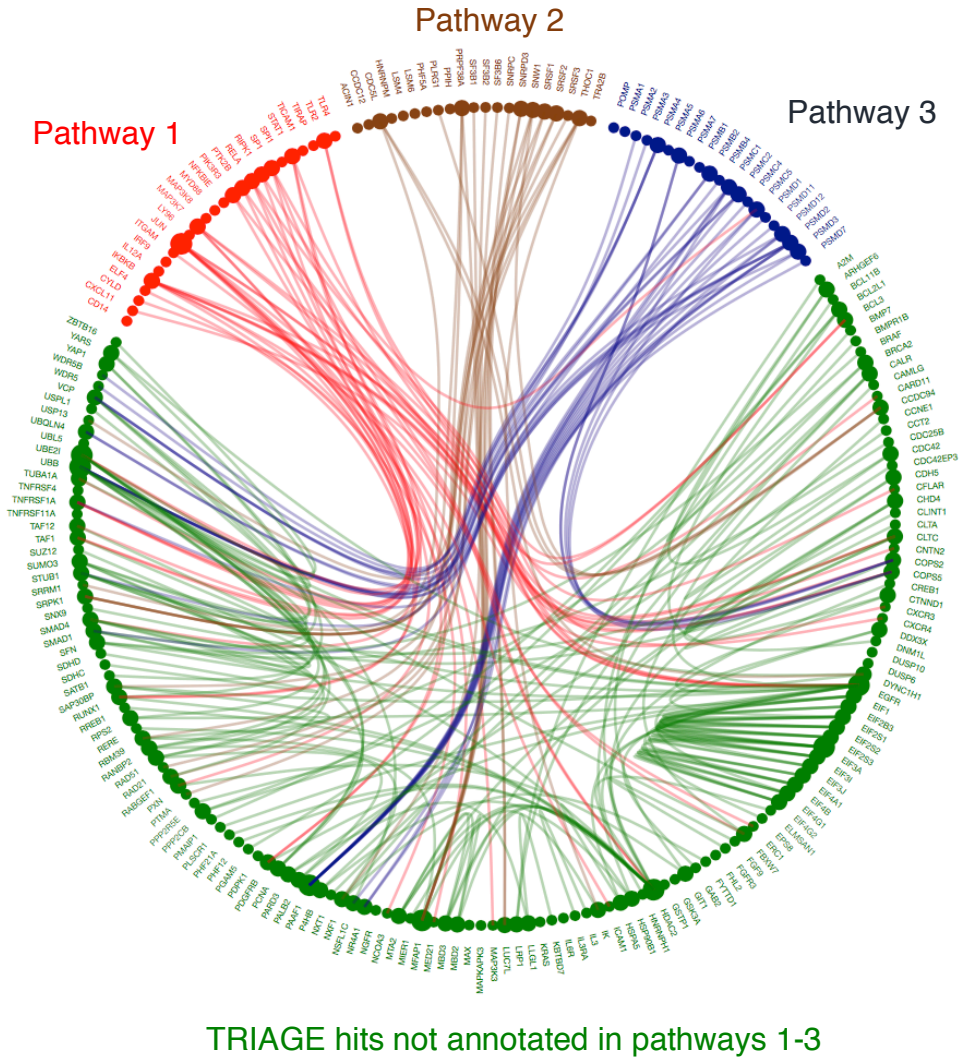


Figure 5.6: **A hierarchical edge bundling for pathway and network visualization.**

TRIAGE hits that are annotates of part of selected pathways are grouped together and highlighted by different colors (red, brown, and blue). TRIAGE hits that are not annotated as part of the selected pathway based on the pathway database but have predicted interactions with pathway genes based on the network databases are added to a new group (green). Interactions are colored based on the group it originates from. Only interactions between the groups and within the group of added genes (green) are shown.

5.16 An interactive version of pathway and gene hierarchical edge bundling

To enable further exploration of the datasets in TRIAGE I developed an interactive version of the integrated pathway and network graph within the platform. The user can select up to three pathways from the list of enriched pathways and TRIAGE will generate an HEB graph of all the TRIAGE hits that are part of the selected pathways as well as all the TRIAGE hits that have predicted interactions with the 'pathway' genes (Figure 5.7). The visual parameters of the graph can be adjusted by the user. For example, hovering with a cursor over a specific gene ("node") highlights all the predicted interactions ("edges") of that gene. Clicking on this gene 'fixes' the interactions, so that the user can then click on one of the predicted interactions to observe the interactions from the second node. A panel at the side of the graph provides information about the interaction, such as the evidence source for the interaction and its confidence score from the STRING database. After clicking through a string of interacting genes the user can click the "Highlight Clicked Pathway" icon and all the genes clicked through in the exploration, together with their predicted interactions, are highlighted (Figure 5.7). The clicked genes, and the pathways they are members of are tabulated in a separate table that can be downloaded with the rest of the analysis at the download tab.

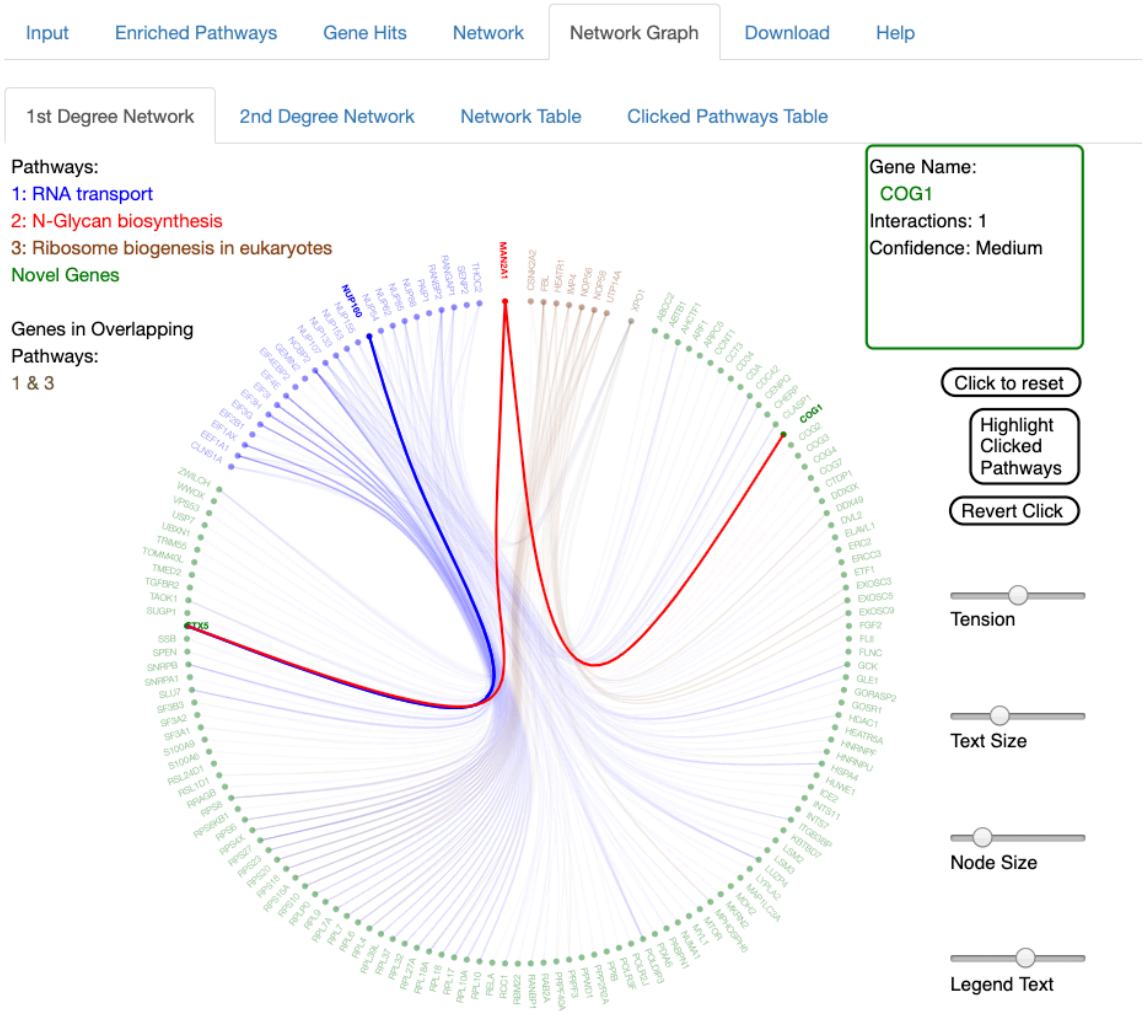


Figure 5.7: **Interactive interface for data exploration of pathway and network connections.**

After selecting up to three enriched pathway an interactive network figure is generated. Showing the results following the selection of the "RNA transport", "Ribosome biogenesis in eukaryotes", and "N-Glycan biosynthesis" pathways in the TRIAGE analysis of the *Brass et al.* study of essential factors for HIV infection. Information about different nodes and edges appear in the window at top right. Aesthetic parameters can be controlled by the sliders on the right. "Highlight Clicked Pathway" highlights the clicked on genes and their reactions, as shown in this figure.

5.17 Steps taken towards data security on triage.niaid.nih.gov

TRIAGE uses a secure encrypted HTTPS connection. To further increase security, when a file is uploaded a new directory is created where the input file and all the subsequently generated analysis files are temporarily saved, ensuring that only the user generating the connection can access the directory with their files. The directory is kept on the server only for the duration of the session (i.e. as long as the user is using the site). Once the sessions ends (i.e. close of browser window or move to a new site) the directory and all its files are removed from the TRIAGE server. This decreases the security risk for the user and ensures that the results are only stored locally after the analysis. Data collected for each session is limited to the country where the request comes from and the time spent using the site (and specific pages). File names, analysis choices, user IDs, and results are neither collected nor stored.

5.18 Summary

The development of a robust hit prioritization method for high-throughput studies of the LPS response in macrophages led me to the identification a novel bioinformatic approach employing iterative analysis of pathway and network databases (TRIAGE). Recognizing that this approach has utility beyond the study I designed, I chose to broaden its access. To ensure accessibility and ease of use, I built a web interface that guides the user through straightforward steps of the TRIAGE analysis. I also built the analysis results interface to enable further exploration of the data via interactive features and integrated information from selected databases. The TRIAGE URL (*triage.niaid.nih.gov)* is currently hosted by the National Institutes of Health (NIH).

6 TRIAGE analysis of LPS screen data identifies a critical and broad regulatory role for spliceosome and proteasome related genes in macrophage activation

6.1 Introduction

Following the development and testing of TRIAGE, I applied the analysis pipeline to the three genome-scale siRNA studies of the macrophage response to LPS I introduced at the beginning of this thesis in chapter 3. The three studies included an siRNA screen using THP1 cells with a luciferase reporter tagged to a TNF-α promoter (Human-TNF-α study), an siRNA screen using RAW G9 cells tracking GFP-tagged p65 (Mouse-NFκB study), and an siRNA screen using RAW G9 cell measuring an mCherry reporter tagged to a TNF-α promoter (Mouse-TNF-α study). (The different reporter cell lines and the design of the three studies are described in detail in Figures 1.6 and 1.7.) All three studies have been previously reported with a number of observations selected for in-depth follow up analysis (John et al., 2018; Ning Li et al., 2017; J. Sun et al., 2017; J. Sun et al., 2016). With the development of TRIAGE, I reanalyzed these datasets to find hits and enrichments beyond the highest scoring genes. In this chapter, I begin by describing segmentation of the data into high and medium confidence hits, and then apply the TRIAGE analysis pipeline to all screens individually. I show that the discordance between the studies that I observed in chapter 3 is largely overcome by TRIAGE analysis (section 6.2). The TRIAGE selected hits also show a more robust enrichment for canonical TLR pathway genes (section 6.3). Looking beyond individual hits, I show how TRIAGE enriches for many immune-related pathways, as expected, as well as two critical biological processes; the proteasome and the spliceosome. Using the integrated pathway and network approach, I also show how the dataset is enriched for hits that provide interaction links between the TLR pathway and the proteasome and spliceosome (section 6.4). As a further validation for a broad role of the proteasome and spliceosome in macrophage activation, I show two experiments using chemical inhibitors of the relevant processes and report on their impact on the LPS

response (section 6.5-6.6). I follow up these studies with an enrichment analysis between an RNA-seq study of the LPS response and the hits from the LPS screens (section 6.7). While alternatively spliced isoforms of TLR pathway genes have been reported before, a critical question to address in this context is to what extent dynamic splicing events occur within the rapid time frame of macrophage activation. I address this question by PCR analysis, using primers flanking alternatively spliced exon boundaries of canonical TLR genes, and demonstrate their alternative engagement within the window of early macrophage activation (section 6.8). By the combined analysis and follow up studies in this chapter I show how the TRIAGE analysis can identify novel regulatory mechanisms by which macrophages may use proteasome-targeted negative regulators to set thresholds for activation, and dynamic alternative splicing to promote and sustain an inflammatory response.

6.2 Analysis by TRIAGE of three LPS screens

To conduct TRIAGE analysis of the three studies of the LPS response, I began by using the normalized datasets I created in chapter 3. The data was already normalized with a cell viability correction (section 3.1). In deciding how to prioritize hits for high or medium confidence, I considered possible additional metrics to the normalized readout scores. siRNAs (and as is being discovered, CRISPR/Cas9 guides studies as well) can have off-target effects (A. L. Jackson & Linsley, 2004; Aimee L. Jackson & Linsley, 2010; Mohr et al., 2010). In these cases, though the perturbation reagents are designed to target a specific gene, they may through weak binding affinities with alternative sequences interfere with the expression of a different gene. These off-target activities can lead to the observed phenotype being erroneously attributed to the incorrect gene. A widely used bioinformatic solution for RNAi is using Common Seed Analysis (CSA) (Marine et al., 2012). Common Seed Analysis looks at all the possible high and low affinity bindings of the siRNA oligonucleotide and compares the readout score of its

target gene versus the scores of its possible off-target partners. CSA provides a statistical score for how likely the observed result is caused by an off-target effect. When prioritizing the hits from the siRNA studies I assigned a cutoff to each study (corresponding to its Z score inflection point) and applied CSA. Genes that were above the cutoff and were not estimated by CSA to be a possible off target effect were assigned as high confidence, genes that were above the cutoff but were flagged as possibly being driven by off target effects were assigned as medium confidence. Similarly, hits that had a score between the set cutoff and a Z score of 1 standard deviation from the mean yet had a low score for possible off-target effects were also considered medium confidence (Figure 6.1A).

Having established the method for ranking the LPS screens into three tiers, I proceeded with TRIAGE analysis. Each of the three screens cycled through 3 to 4 iterations of hit prioritization until the analysis converged on a final set of hits. Though the three studies had different distributions, the analysis by TRIAGE found similar sized sets for the Human-TNFα and the Mouse-NFκB studies. The mouse-TNFα study started with a very small set of hits with analysis by TRIAGE broadening it to include more lower scoring hits (Figure 6.1B).

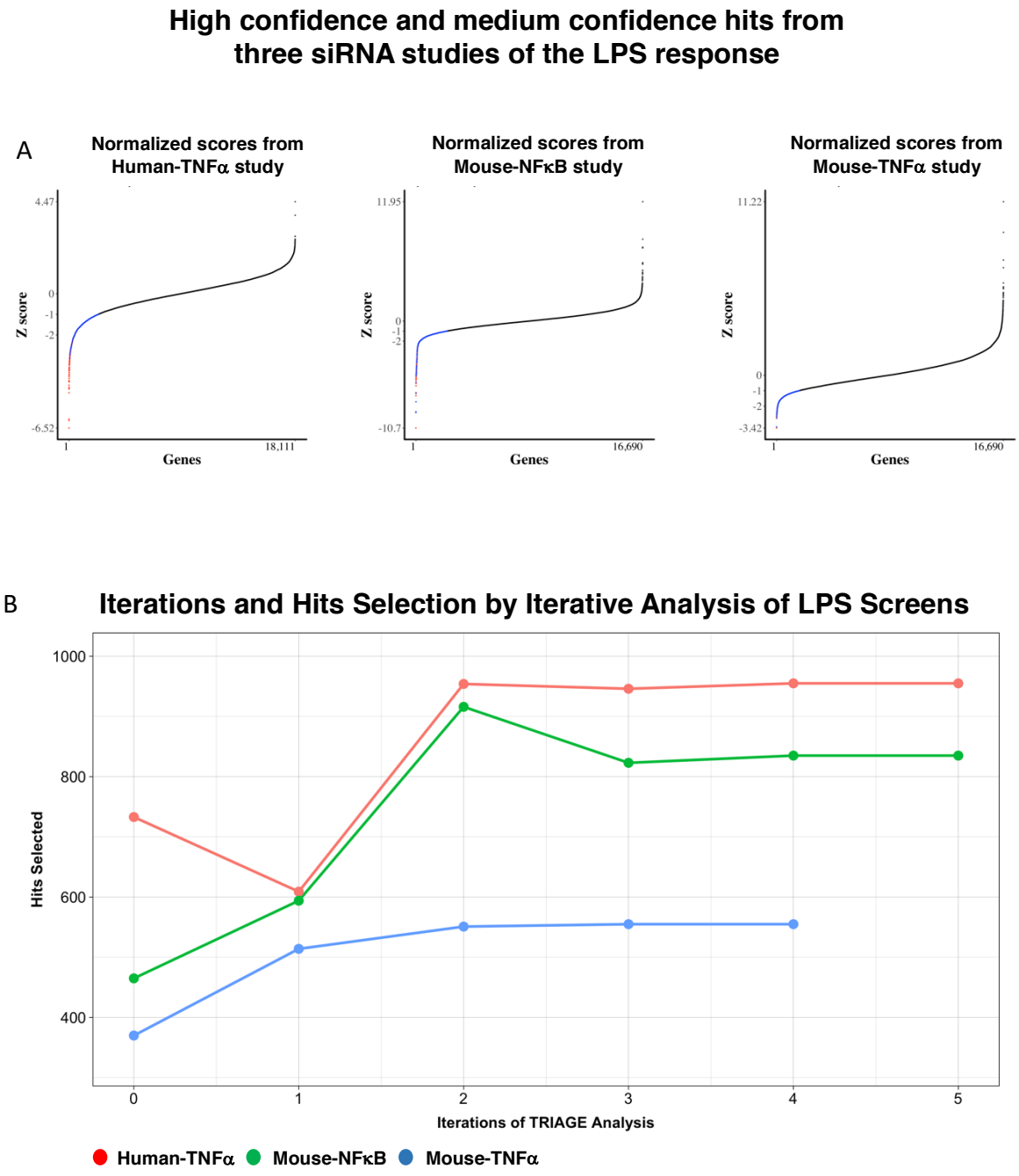


Figure 6.1: **TRIAGE analysis for three siRNA studies of the response to LPS.**

(A) Normalized distribution of the readout scores for the human TNF-α study (left), mouse NF-κB study (center), and the mouse TNF-α study (right). High-confidence assigned hits are in red, medium confidence in blue. (B) Iterations of TRIAGE analysis of the three LPS response studies.

As an initial test of confidence in the selected hits I compared the overlap across the three studies. As I have shown in chapter 3, overlap between the highest scoring hits from the three studies is quite low. Hit selection by TRIAGE, however, identifies a very strong shared enrichment between the studies. The overlap across the studies also have a strong statistical significance (Figure 6.2) suggesting a robust true positive rate.

6.3 Enrichment for canonical TLR pathway genes in the three screens of the LPS response are shared and divergent

Following analysis by TRIAGE I used the Ingenuity Pathway Analysis (IPA) program to test the selected hit sets for enrichment of established TLR pathway components. In contrast to the enrichments found in the high scoring hits in chapter 3 (Figures 3.2-3.4), hits selected by TRIAGE showed enrichment for canonical TLR pathway components across all the major signal transduction junctures (Figure 6.3-6.5). Critically, the three studies show enrichment in all three "modules" of the TLR pathway (encoder, transmission, and decoder, described in the introduction section 1.5.2, Figure 1.5). The encoder module (TIRAP, TICAM1, and TOLLIP, in addition to CD14, MyD88, and TLR4), the transmission module (TAB2, TAK1, TRAF6, IKKα), and the decoder module (NFκB, c-Fos, MAPK, p38). Of interest, TRAF6 only came up as critical in the Mouse-NFκB screen, suggesting that the TNFα response is less sensitive to TRAF6 perturbation than NFκB. TAB2 and TAK1 were only hits in the two TNFα studies, suggesting that the MAPK p38 and Jnk responses downstream of TAK1 are critical to support the TNFα readout (Figure 6.3 and Figure 6.5).

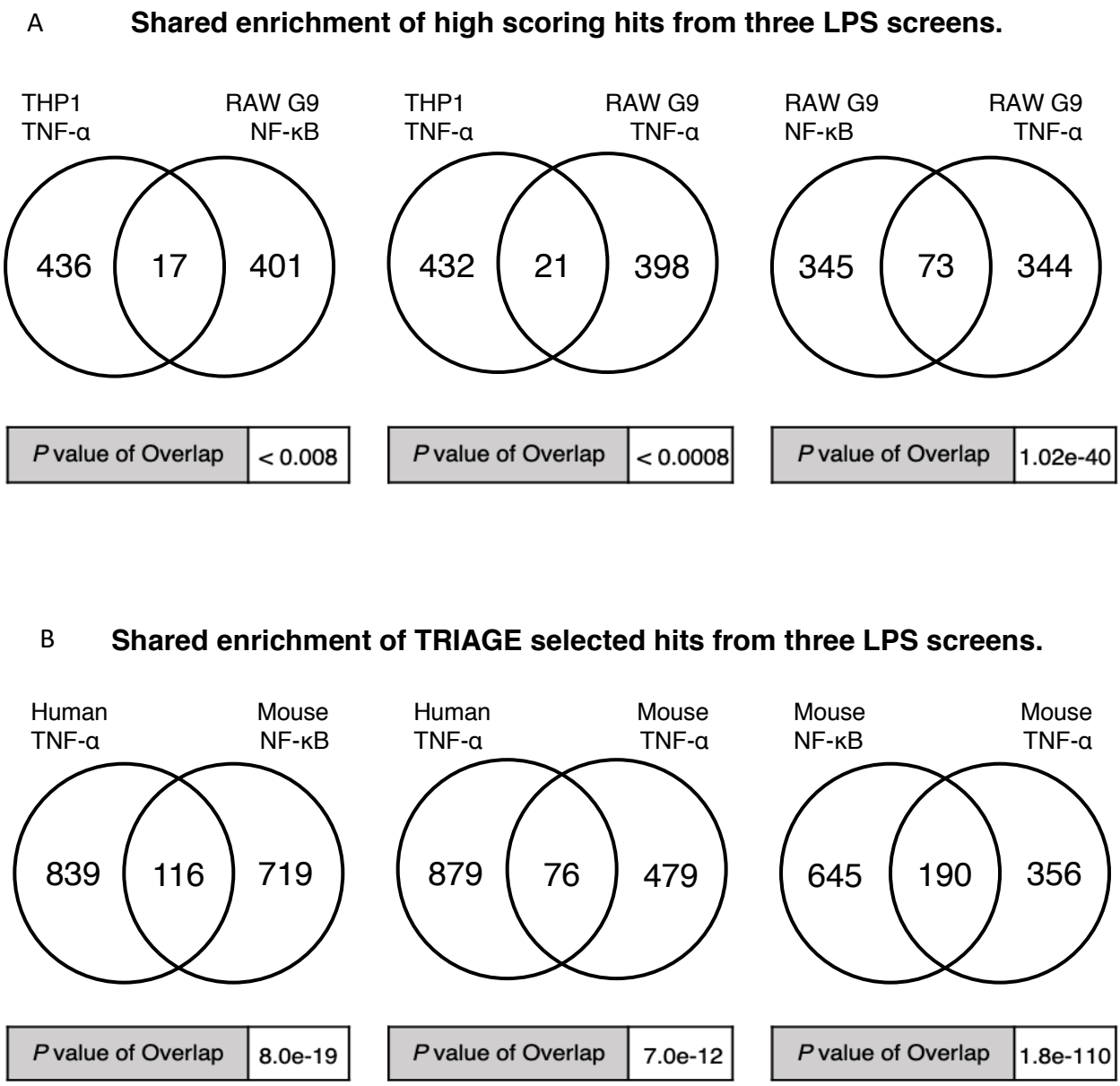
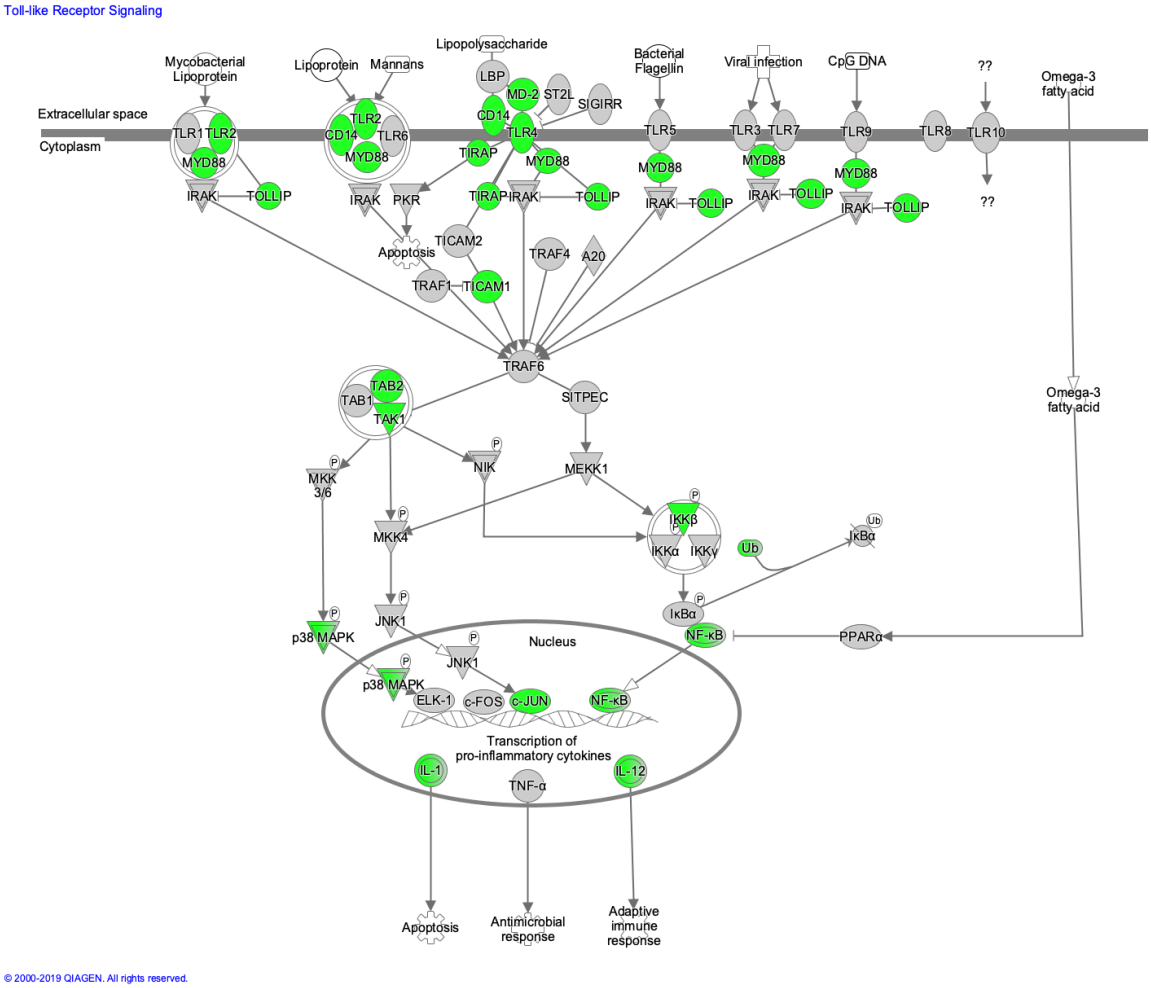


Figure 6.2: **Shared enrichment of TRIAGE selected hits from three siRNA studies of the response to LPS.**

(A) Number of shared hits and significance of overlap between the highest scoring 2.5% of hits from three screens of the macrophage response to LPS. (B) Number of shared hits and significance of overlap between hits selected by TRIAGE from the three screens of the macrophage response to LPS. *p* value of overlap is calculated by the hypergeometric test.

Enrichment of TLR pathway genes in TRIAGE selected hits from the THP1 TNF-α study

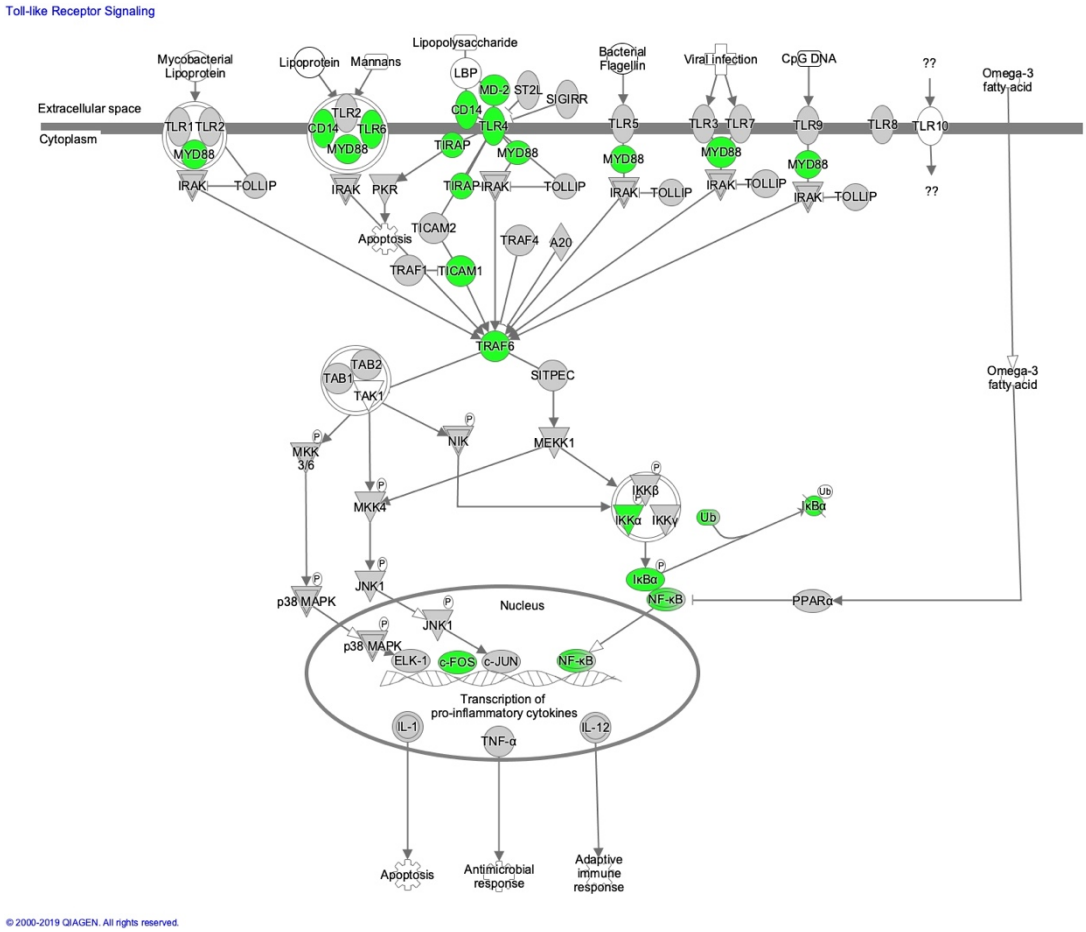


Putative Positive Regulators

Figure 6.3: **Enrichment of canonical toll-like receptor pathways genes in TRIAGE selected hits from THP1 TNF-α genome-wide studies.**

Overlaying TRIAGE selected hits from the THP1-dual luciferase study over the canonical pathway map curated by Ingenuity Pathway Analysis (IPA, Qiagen)

Enrichment of TLR pathway genes in TRIAGE selected hits from the Raw G9 NF- κB study

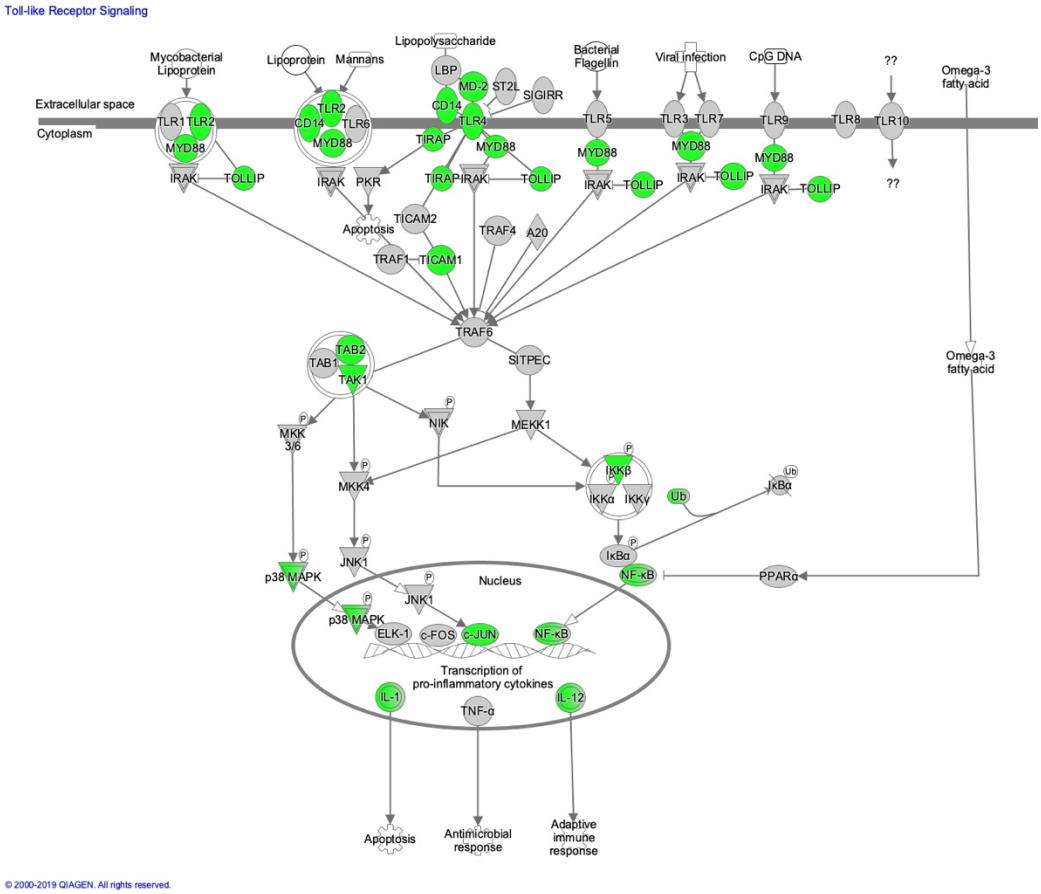


Putative Positive Regulators

Figure 6.4: **Enrichment of canonical toll-like receptor pathways genes in TRIAGE selected hits from Raw G9 NF- κB genome-wide studies.**

Overlaying TRIAGE selected hits from the from the RAW G9 – GFP tagged NF-κB study over the canonical pathway map curated by Ingenuity Pathway Analysis (IPA, Qiagen)

Enrichment of TLR pathway genes in TRIAGE selected hits from the Raw G9 TNF-α study



Putative Positive Regulators

Figure 6.5: **Enrichment of canonical toll-like receptor pathways genes in TRIAGE selected hits from Raw G9 TNF-α genome-wide studies.**

Overlaying TRIAGE selected hits from the from the RAW G9 – mCherry tagged TNF-α study over the canonical pathway map curated by Ingenuity Pathway Analysis (IPA, Qiagen)

6.4 Immune, spliceosome, and proteasome pathways are critically enriched in the three studies of LPS response

For pathway analysis of the three LPS screens, I prioritized enrichments that were found in at least two out of the three studies. In addition to the *p*-value threshold, I added a False Discovery Rate (FDR) test and set the threshold at FDR = 0.1. The pathways identified included, as expected, the TLR pathway itself, integral sub-components of the TLR response (including NFκB and MAPK) and many other immune related pathways in the top scoring and shared enrichments (Figure 6.6). Two pathways that are not primarily immune related but appear as strong enrichments in the study are the proteasome and spliceosome (Figure 6.6). Though the proteasome is to be expected as critical for TLR signaling, given the necessity of proteasomal degradation for removing negative inhibitors (as I described in the introduction, section 1.3.1), the strong enrichment of multiple proteasomal factors suggested the possibility that there is an even broader requirement for the proteasome in TLR activation. As I have shown in the previous section of the TLR enrichment of the three screens, hits from the TNFα screens indicate a signaling pathway independent of NFκB. The strong enrichment of proteasome factors in the Mouse-NFκB and the Human-TNFα screens, suggests that the function of the proteasome in enabling macrophage activation goes beyond relieving inhibition of NFκB translocation. There are similar considerations with the enrichment of the spliceosome pathway. Though the core components of the spliceosome are critical for basic cell function, the enrichment of splicing related hits even after applying a critical cell viability correction (section 2.3.3), suggest a possible function beyond cell homeostatic requirements that is specific to TLR signaling. To further highlight these enrichments, I followed up the pathway analysis with an integrated pathway and network analysis to identify the hits from the screen that aren't annotated as part of the TLR, proteasome, or spliceosome pathways, yet have predicted interactions with gene members of at least one of those pathway sets. The enrichment

found further representation of splicing and proteasome related genes, many with known interaction to TLR components (Figure 6.7). In all three LPS studies, HNRNPH1 and NXF1 were identified as hits. Though not formally annotated in the spliceosome pathway by KEGG, their enrichment suggests a path for the engagement of its main components.

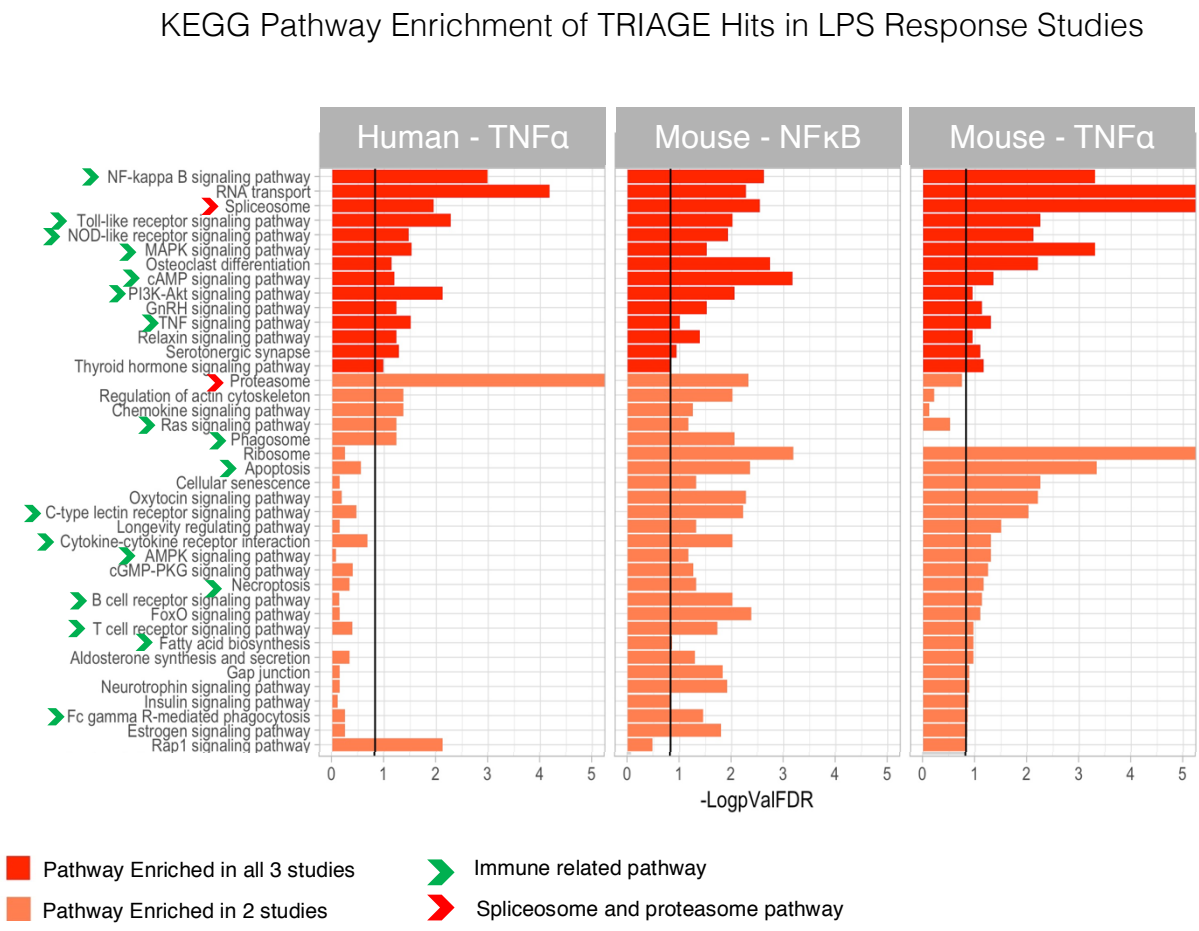


Figure 6.6: **Pathway enrichment by TRIAGE of three studies of LPS response.**

KEGG pathway enrichment analysis of the human TNF-α study (left), mouse NF-κB study (center), and mouse TNF- α study (left). Only showing shared enrichments. Red bars indicates pathways enriched in all three studies. Orange indicates enriched in two out of three studies. Significance indicates a cutoff of FDR < 0.1, Figure truncated at –Log(pValFDR) = 5.

Network analysis of TRIAGE selected hits linking the TLR pathway with mRNA splicing and protein degradation

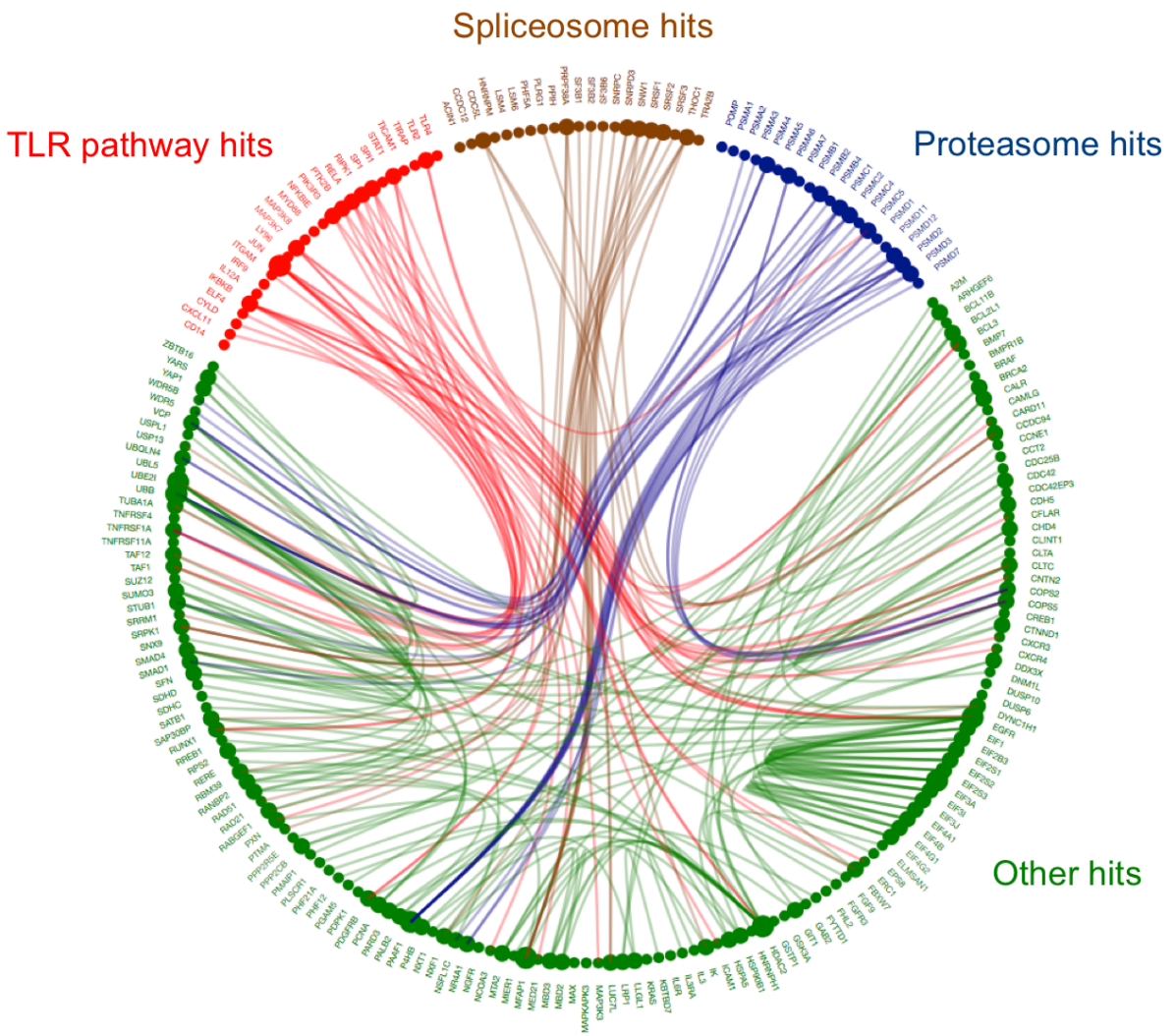


Figure 6.7: **Network and pathway integrated analysis of TLR, spliceosome, and proteasome related hits.**

Integrated pathway and network analysis of the TRIAGE selected hits from the human- TNF- α study. The network shows hits from the study in the TLR pathway (red), spliceosome pathway (brown), proteasome pathway (blue), and hits not annotated as part of the three pathway, yet have predicted network interaction with hits in the three pathways.

6.5 Inhibition by MG132 shows proteasome degradation is critical for signaling in multiple branches of the TLR pathway

To investigate further the possible basis for the strong enrichment of proteasome components in the LPS screens, we investigated the effects of proteasome inhibition on LPS-induced signaling in macrophages. MG132 (carbobenzoxy-Leu-Leu-leucinal) is a proteasome inhibitor that has been used previously in macrophage cells (Ortiz-Lazareno et al., 2008; Perry et al., 2010). For this assay I worked with Dr. Jing Sun at the National Institutes of Health (NIH), and we sought to determine whether the proteasome is required for LPS signaling beyond the known requirement for degradation of IκB and p105 (Ciechanover et al., 2001; Coux & Goldberg, 1998). We identified a juncture in the downstream signaling of LPS that is independent of these two factors. p38 MAPK is a signaling pathway independent of NFκB and it is also a key contributor to the LPS-induced TNF response (Gottschalk et al., 2019; J. Sun et al., 2016). We chose this juncture in addition to two factors from the NFκB and p105 signaling pathway, IκB and Erk respectively. Using antibodies for IκB, phospho-Erk, and phospho-p38, we find that the LPS-induced degradation of IκB as well as the activating phosphorylation of Erk and p38 are all substantially diminished in the presence of MG132 (Figure 6.8A-C). This suggests that proteasomal function is critical in the activation of all the signaling branches tested (NFκB, Erk MAPK and p38 MAPK), and further indicates that the strong enrichment for proteasome components in all three screens could be due to a pervasive requirement for protein degradation to activate numerous signaling events downstream of TLR4 activation.

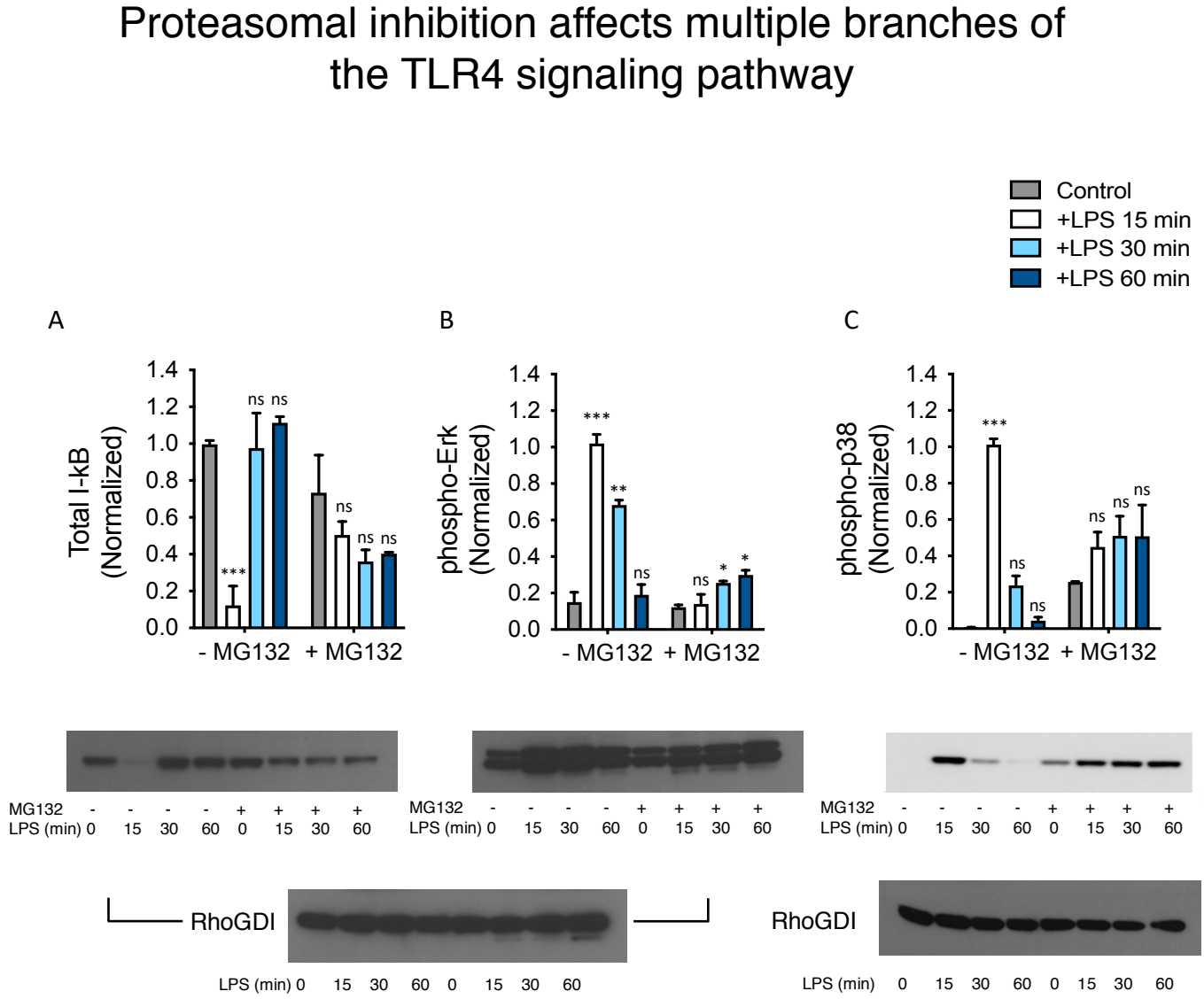


Figure 6.8: **Changes in phosphorylation of TLR effectors following proteasomal inhibition.**

(A) Western blot of Total IκB following 15 minutes, 30 minutes, and 60 minutes of LPS treatment shows only small changes in Total IκB following proteasomal inhibition. (B) Levels phospo-Erk (a factor downstream of p105) following LPS treatment with and without MG132 inhibition of the proteasome. (C) Levels of phopho-p38, which signals independently of the NFκB pathway, in the absence and presence of proteasomal inhibition by MG132. Bar graphs show the results of two or more experiments. Western blot gels show representative samples. * = $p < 0.05$, ** = $p < 0.01$, *** = $p < 0.001$, multiple t test.

6.6 Inhibition of splicing by Madrasin blocks transcriptional response to LPS

To first test in an overarching manner whether engagement of the spliceosome is essential for the activation of LPS signaling, I used a chemical inhibitor of splicing to observe its impact on signal transduction. The current model of signal transduction in TLR4 involves the recruitment, conformational alteration, degradation, and translocation of proteins that lead to pro-inflammatory transcription and cytokine expression. While it can be assumed that the expression of pro-inflammatory cytokines would require the essential machinery of mRNA splicing and translation, it would be a novel finding if the signaling upstream of pro-inflammatory transcription actively depends on the engagement of the spliceosome to facilitate and maintain signal transduction.

To test this hypothesis, I utilized the RAW G9 reporter cell line from the LPS screen together with a chemical inhibitor of splicing. The RAW G9 cell line contains an mCherry reporter that is tagged to the single exon promoter of TNF-α (Figure 1.4C). The expression of the single exon promoter does not require any splicing to activate the reporter activity, therefore this system can be used as a way to measure splicing activity required upstream of the initiation of TNFα transcription. The mCherry reporter is tagged to a PEST sequence which has an approximate 1 hour half-life (Sung et al., 2014). The short half-life provides a dynamic readout of continued activation of the TNF-α promoter from the measured fluorescence. Several chemical inhibitors of splicing have been identified (Effenberger et al., 2017), however, many are derived from bacteria and sources that lead to trace levels of contaminants that can be recognized by PRRs and activate macrophages. Madrasin (2-((7methoxy-4-methylquinazolin-2-yl)amino)-5,6-dimethylpyrimidin-4(3*H*)-one RNAsplicing inhibitor) is a synthetic inhibitor of splicing (Pawellek et al., 2014) (Figure 6.9A) which makes it a better candidate for use in immune cells such as macrophages. Madrasin targets the A complex of the spliceosome, blocking the SN2 reaction between two proximal exons junctions. At the outset, I first tested

that the treatment by Madrasin did not affect the basal expression of the TNF promoter-driven mCherry reporter (Figure 6.9B). I then did a dose response assay of 4-hour pretreatment with increasing concentrations of Madrasin followed by treatment with LPS. Measuring mCherry expression in the RAW G9 cells with increasing dosage of Madrasin showed a clear dose response curve of inhibition (Figure 6.9C). I paired this assay with a cell viability assay measuring the levels of Lactate Dehydrogenase (LDH) compared to a cytotoxic control. Pre-treatment by Madrasin did not significantly increase cell death, reducing the likelihood that the observed effect on the TNFα response is driven by a decrease in viable cells (Figure 6.9D). I followed up these assays with a time course assay. Pre-treatment with Madrasin completely attenuated the LPS response, as measured by the *Tnf* promoter-driven mCherry expression (Figure 6.10). These results further support our finding from the enrichment analysis of the LPS studies that the spliceosome is required for LPS signaling upstream of the transcriptional response.

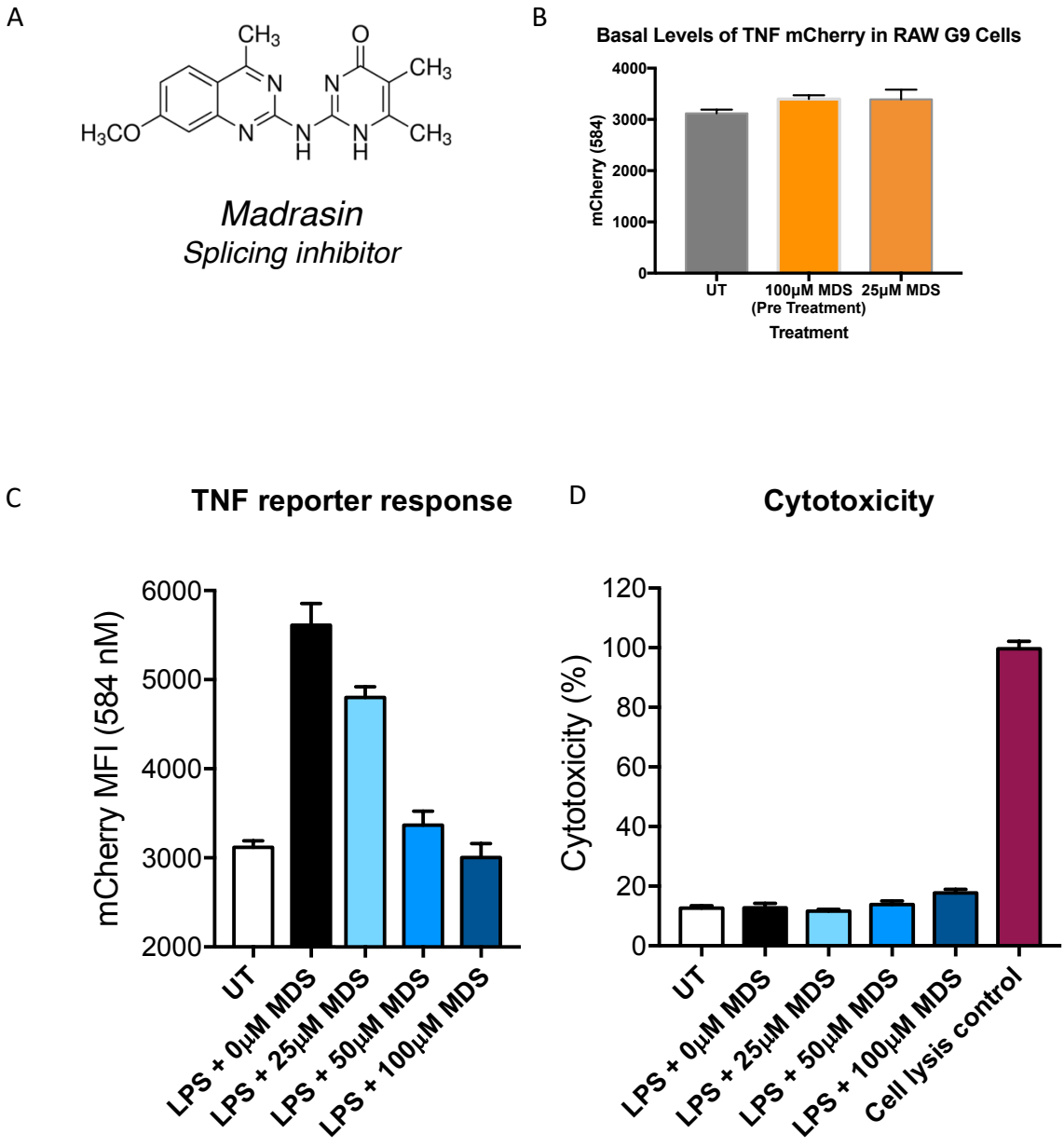


Figure 6.9: **Dose response and cytotoxicity of splicing inhibition and TNF reporter response.**

(A) Structure of 2-((7methoxy-4-methylquinazolin-2-yl)amino)-5,6-dimethylpyrimidin-4(3H)-one RNAsplicing inhibitor (Madarasin) (from Sigma-Aldrich). (B) Measurement of basal expression of the mCherry TNF –α reporter in RAW G9 cells treated with pretreated with Madrasin (MDS). (C) Dosage response of Madrasin pretreatment and LPS (10ng/mL) shows staggered impact on mCherry TNF- α reporter. (D) Cytotoxicity assay measuring LDH release as compared to cell lysis control. All experiments were done in triplicates, with error bars representing deviation of the samples. Experiments were repeated two times, data shown are from representative sample.

Single Exon TNFα-mCherry reporter response to LPS is Ablated by Spliceosome Inhibition

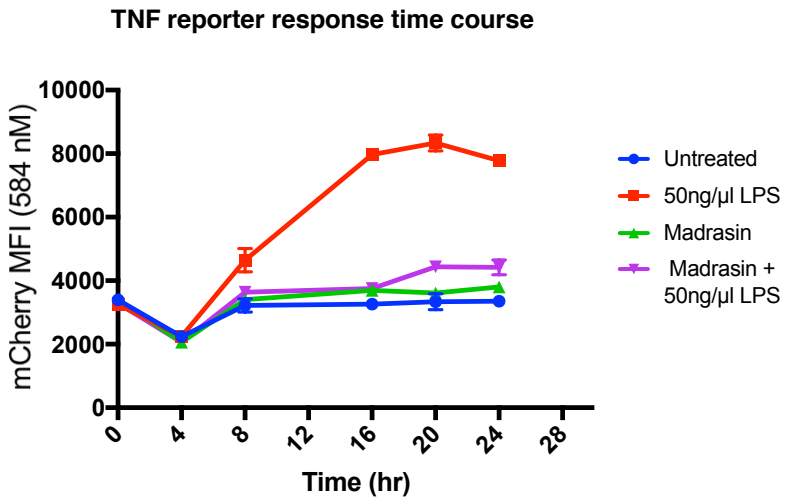


Figure 6.10: **Time course assay of splicing inhibition effect on LPS response.**

Time course assay of RAW G9 cells pretreated for 4 hours with 100uM Madasin showing ablation of the LPS response by splicing inhibition. TNF –α expression was measured by expression of mCherry tagged reporter. Untreated cells were given similar concentrations of DMSO as treated cells. All timepoints were done in triplicates, error bars represent deviation across replicates. Experiment was repeated two times, data shown is from a representative sample.

6.7 LPS screen hits are enriched for predicted interactions with alternatively spliced genes of the TLR4 pathway

To consider alternative splicing as a regulatory mechanism for the activation of macrophages, it is essential to also determine whether effectors of the TLR pathway are alternatively spliced within the time frame of the LPS response. As a first step evaluation, I analyzed an RNA-seq dataset generated by Dr. Sinu John and Dr. Mani Narayanan at the NIH. The RNA-seq dataset used RNA from wildtype THP1 cells as well as THP1 cells with the expression of UBL5 knocked down by shRNA (shUBL5). UBL5 is a factor involved in pre-mRNA splicing (Oka et al., 2014), and it was selected by Dr. John for further analysis as one of the strongest hits in the THP1 LPS screen dataset. The dataset measured differential exon usage following stimulation by LPS for 0.5, 1, 2, and 4 hours. I conducted an analysis of their dataset to look for TLR pathway genes that have changes in exon usage following LPS treatment, and found that many canonical genes from the TLR pathway show relative changes in exon levels in wildtype THP1 that were diminished in shUBL5 cells (Figure 6.11). These results suggest that many effectors of the TLR pathway are alternatively spliced in response to treatment by LPS and that some of the changes depend on a specific splicing factor.

To further illustrate the enrichment of potential regulatory splicing factors, I analyzed by network analysis the hits from our Human-TNFα screen together with the TLR pathway genes that showed differential exon usage following treatment of LPS (Figure 6.12). I used the splicing network categories described by Papasaikas which separates the splicing network into core, hnRNPs, SR proteins, and splicing support to categorize the spliceosome related hits found in the screen (Papasaikas et al., 2015). The network analysis found enrichment in the Human-TNFα screen for all four groups of splicing factors as well as predicted interaction between hits and TLR genes that show differential exon usage in the RNA-seq dataset (Figure 6.11). These results further suggest that both the measurement of alternative splicing in TLR

factors and the engagement of the splicing network could shed insight into how the TLR pathway utilizes alternative splicing in the transition to inflammatory activation.

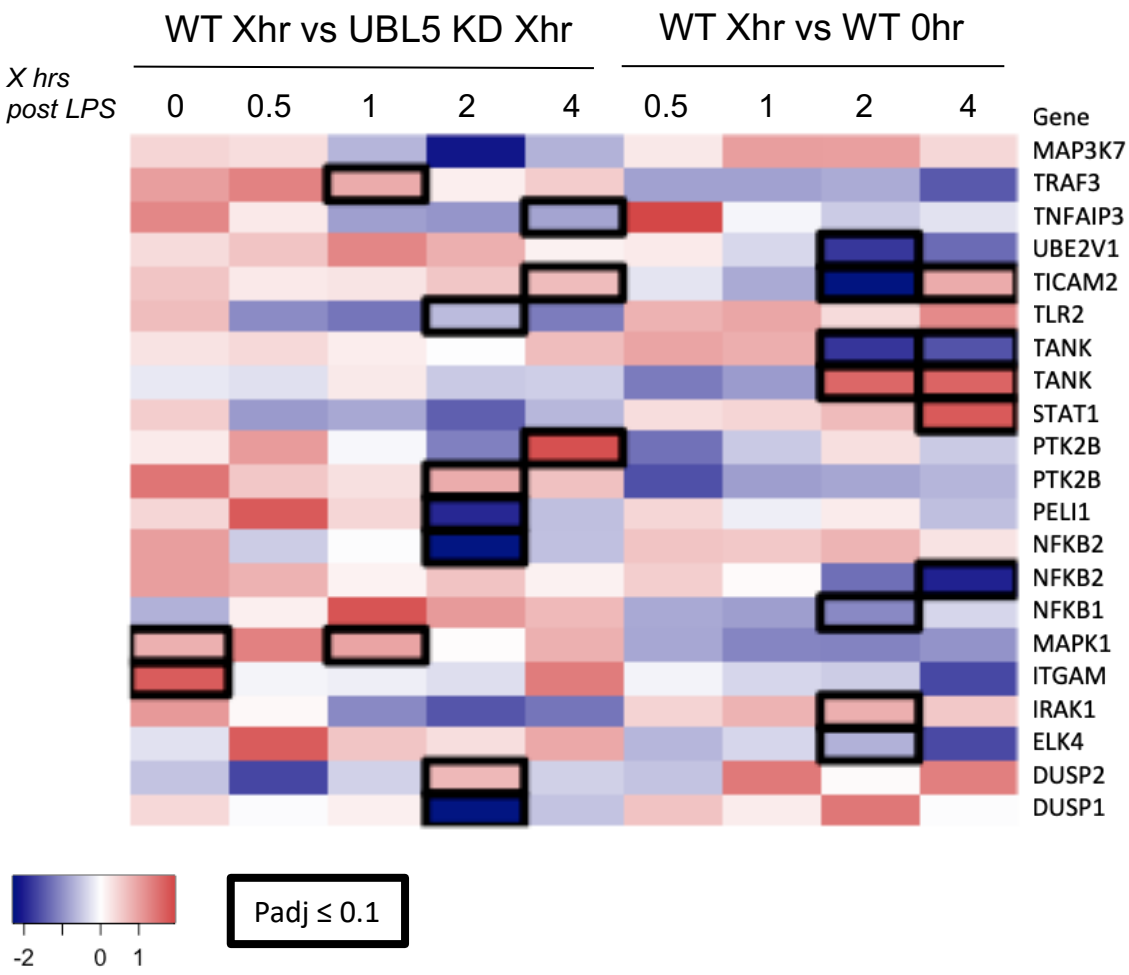


Figure 6.11: **Differential exon usage by TLR genes in WT and UBL5 knockdown cells, following treatment by LPS.**

Relative exon usage was measured by comparing reads from WT cells vs UBL5 KD cells at the same time point after LPS treatment (left panel) or WT at a specific time after LPS treatment as compared to untreated WT cells. Exon usage was calculated by DexSeq. Fold change measurements that have an FDR of 0.1 or less are highlighted by a black border. TLR genes with a fold change that crossed the FDR threshold of 0.1 in at least one time point were included.

Network analysis of TLR genes with differential exon usage and TRIAGE selected hits linking core and supporting spliceosome genes

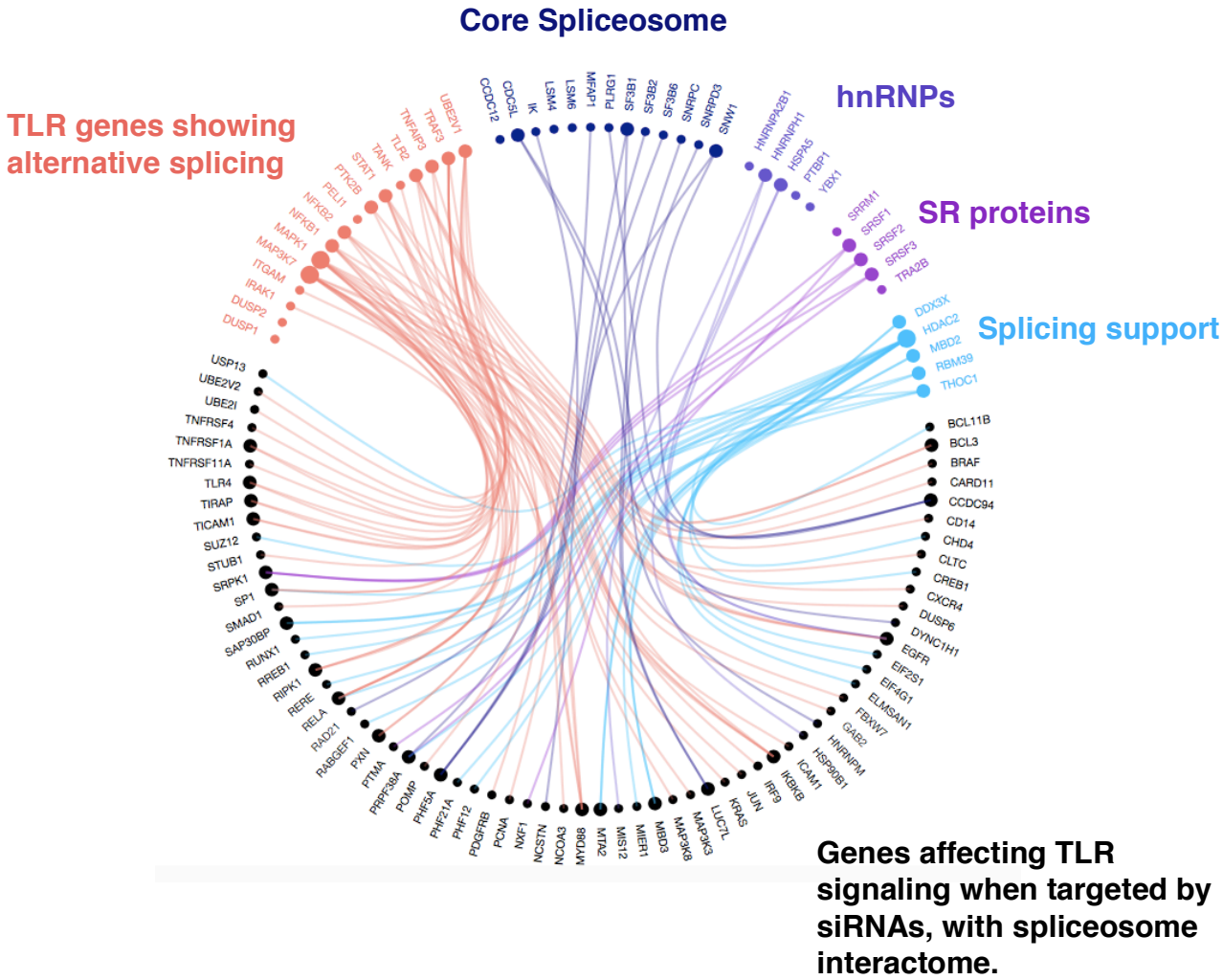


Figure 6.12: **Network analysis of TLR factors showing differential exon usage by RNA-seq and spliceosome network hits from LPS response study.**

Network analysis using curated interactions from the STRING database. (Evidence source = Experimental and score ≥ 700). The TLR genes were selected based differential exon usage by RNA-seq. The splicing factors were selected as hits from the Human TNF –α study. The splicing factors were grouped based on the categories outlined by *Papasaikas, 2015.*

6.8 Examples of alternative splicing in response to LPS treatment

Alternative splicing acting as a regulatory mechanism for TLR activation relies on the possibility that the splice variant repertoire of TLR pathway components changes dynamically soon after LPS stimulation. To further interrogate whether such changes are observed, I selected two TLR factors from the RNA-seq dataset that have known splice variants. IRAK1 has reported variants that lack exon 11 (Rao et al., 2005). TMED7-TICAM2 is a gene whose alternative splice variants lead to the expression TMED7, TRAM, and TAG genes (Doyle et al., 2012). To assess whether these events occur within the timeframe of the LPS response, I designed sets of primers that span the alternatively spliced regions in the genes and sets of primers that sit outside of it. I then ran a polymerase chain reaction (PCR) using RNA from THP1 cells treated with LPS for 2 or 4 hours, alongside RNA from untreated cells. Running the PCR products from the IRAK1 set of primers on a gel showed that the sets of primers that bound outside of alternatively spliced region stayed consistent from 0 to 4 hours of LPS treatment. The PCR product from the reaction using the set primers that bound to the alternatively spliced region decreased with 2 and 4 hours of LPS treatment (Figure 6.13).

Results from the TMED7-TICAM2 reactions show an interesting discrepancy. Reactions with primers that bound to either the TMED7 gene exclusively or the TICAM2 gene exclusively show a consistent PCR result across the hours of treatment with LPS. The set of primers that bound across the span of TMED7-TICAM2, however, showed an observable decrease following treatment with LPS (Figure 6.14). Suggesting that the mRNA of the TMED7-TICAM2 gene is transcribed and spliced differently in response to LPS treatment, and that its change is observable in the early stages of LPS stimulation.

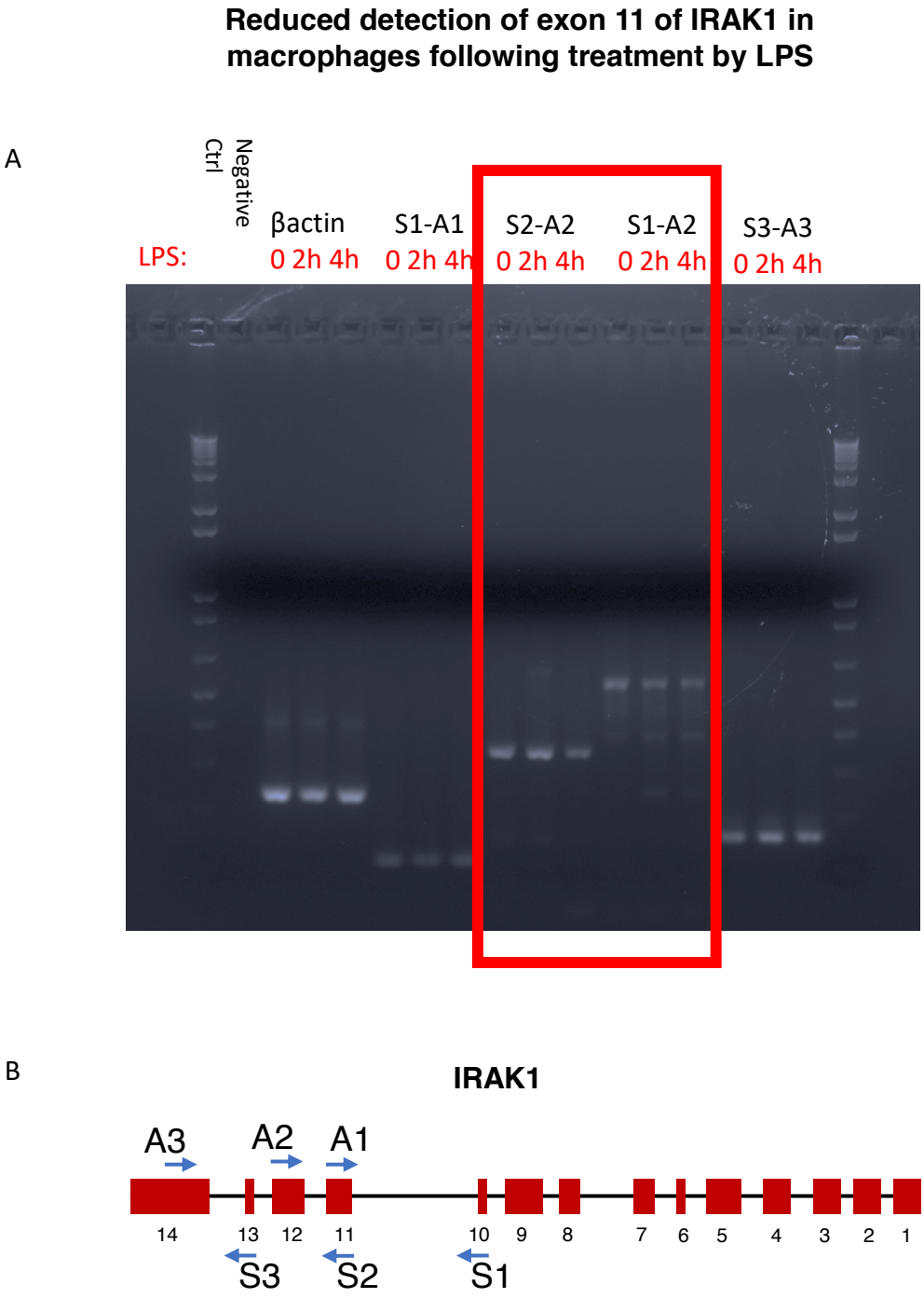


Figure 6.13: **cDNA amplification by primer pairs of IRAK1 following treatment by LPS.**

(A) Primers were designed to align with different exons of IRAK1. Primer sets including exon 11-exon 12 (S2-A2, S1-A2) show a decrease following 4 hours of LPS treatments. Primer sets outside of the exon 11-exon 12 range (S3-A3) shows a slight increase with treatment by LPS. (B) A schematic of IRAK1 exons (based on the NM_001569 NCBI sequence) and the primer sequence placements.

Reduced detection of full length TMED7-TICAM2 gene after treatment of LPS

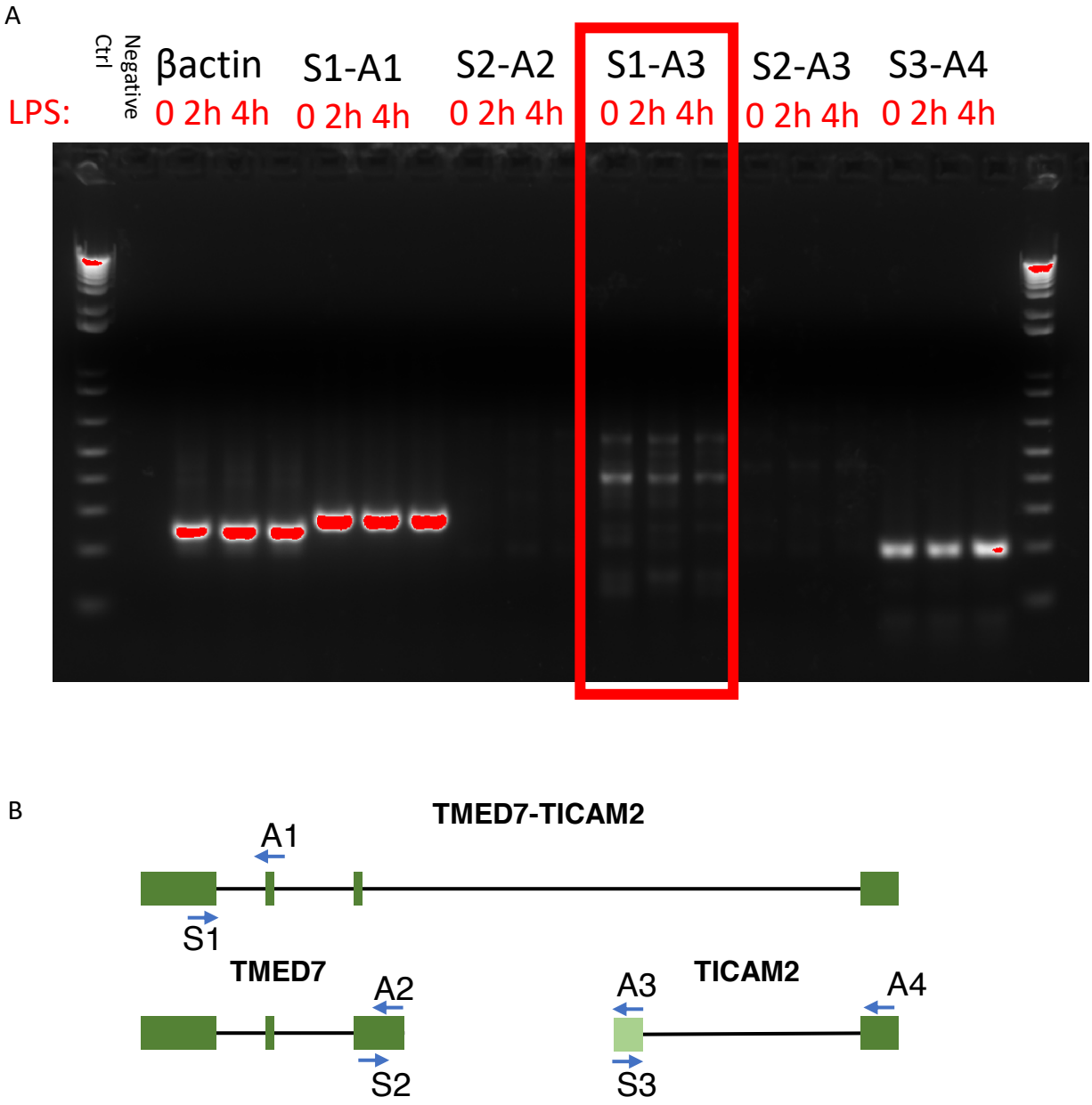


Figure 6.14: **cDNA amplification of TMED7-TICAM2 gene following treatment by LPS.**

(A) Primers were designed to align with different exons spanning different segments of the TMED7-TICAM2 gene. Primer sets that span only the range of TMED7 (S1-A1) or only the range of TICAM2 (S3-A4) shows consistent expression following LPS treatment. The primer set spanning the combined gene transcript (S1-A3) shows a consistent decrease following treatment by LPS. (B) A schematic of TMED7-TICAM2 cDNA sequences (based on the NCBI sequences: NM_001164469 (TMED7-TICAM2), NM_181836.6 (TMED7), NM_021649.7 (TICAM2)) with the primer sequence placements.

6.9 Summary

Analysis by TRIAGE identified robust shared enrichment between the three screens of the macrophage response to LPS. An analysis of the pathway enrichments in the selected datasets show that, in addition to the immune related factors, pre-mRNA splicing and proteasomal factors have a key role in facilitating TLR activation in response to LPS treatment. I followed up this observation with assays to further investigate the roles of the proteasome and mRNA splicing in response to LPS. Treating macrophages with a proteasome inhibitor showed a broad dependency of multiple branches of the TLR pathway on stimulus-induced proteasomal degradation to permit signal transduction. Further validation and studies are still needed, however, to establish how many branches of the pathway it effects and to determine if there are additional unappreciated targets of proteasomal degradation in the LPS signaling cascade. Treating macrophages with an inhibitor of splicing also showed that upstream of cytokine transcription, an active spliceosome is required to support and sustain the LPS response. Analysis of RNA-seq data and follow up by PCR shows that alternative splicing of TLR factors occur within the rapid time frame of innate immune activation. These combined results in alternative splicing and spliceosome inhibition suggest that the engagement of the spliceosome is a critical mechanism in the activation of the inflammatory response. How these dynamic changes occur and what are the impacts on signaling of alternatively spliced factors of the TLR pathway remain to be explored.

7 Short read and long read RNA-seq characterizing alternative splicing in response to LPS

7.1 Introduction

With the identification of a regulatory role for the spliceosome (chapter 6), a critical question to address is whether the macrophage transcriptome is dynamically changed in response to a challenge by LPS. I have shown isolated examples of such changes in chapter 6 (section 6.8), yet to address how such changes are driven by LPS, and what role they could potentially play in the activation of macrophages, requires a broader approach. The next generation sequencing (NGS) method RNA sequencing (RNA-seq) provides a platform where the characterization of cellular RNA can be done on an omic scale (Z. Wang et al., 2009). Advances in bioinformatic analysis pipelines have further expanded the capacity of these analyses to discriminate between different kinds of alternative splicing events (Shen et al., 2014). Third generation sequencing approaches make it further possible to read transcripts end-to-end, eliminating the need to rely on a reference genome to reconstruct transcripts thereby enabling the identification of completely novel isoform variants (Eid et al., 2009). In this chapter I describe the application of these approaches to compare the transcriptome of unstimulated macrophages to the transcriptome of macrophages treated with LPS. For this analysis I utilize the RNA of macrophages differentiated from primary and immortalized monocytes (section 7.2). Using Illumina Iso-seq analysis I show how differential isoform expression is induced by treatment with LPS (section 7.3). Utilizing the rMATS analysis pipeline I show how different forms of alternative splicing are similarly induced in the different cells sequenced (section 7.4). Despite modest overlap on the gene level, I show how the enrichment for TLR signaling pathways in LPS induced alternative splicing is shared in human and mouse (section 7.5). I also show how long read RNA-seq by SMRT Pacific Biosciences more thoroughly identifies the variant changes driven by LPS stimulation (section 7.6). I was assisted in the work of this chapter by the Center for Cancer Research Sequencing Facility (CCR-SF) at Frederick National

Laboratory for Cancer Research (FNLCR) and the NIAID Collaborative Bioinformatics Resource (NCBR) group.

7.2 RNA extraction of primary and immortalized macrophages treated with LPS

To increase the likelihood that the variants identified illuminate the LPS response in human health and disease, and to also have a pipeline for experimental validation in mouse models and cell lines, I ran the sequencing using macrophages derived from human and mouse cells. I included macrophages derived from human peripheral blood monocytes from healthy donors, macrophages differentiated from mouse bone marrow, and macrophages differentiated from the human THP1 cell line. I also included macrophages derived from THP1 cells which had the splicing factor UBL5 knocked down (described in section 6.7 and section 2.5.4). I differentiated the different sourced monocytes using the relevant differentiation factors (GM-CSF for human PBMCs, MCSF for mouse BMDMs, and PMA for THP1s). Following differentiation, I treated a subset of cells with LPS for four hours while leaving a subset of cells untreated. I prepared three independent biological replicates for every condition and extracted total RNA from the cells to be sequenced. A schematic of the experimental design is in Figure 7.1. I also tested to confirm that the LPS treatment worked by measuring the relative induction of TNF-α mRNA by qPCR in the treated versus untreated cells (Figure 7.2).

Sample preparation for RNA-seq of the LPS induced transcriptome

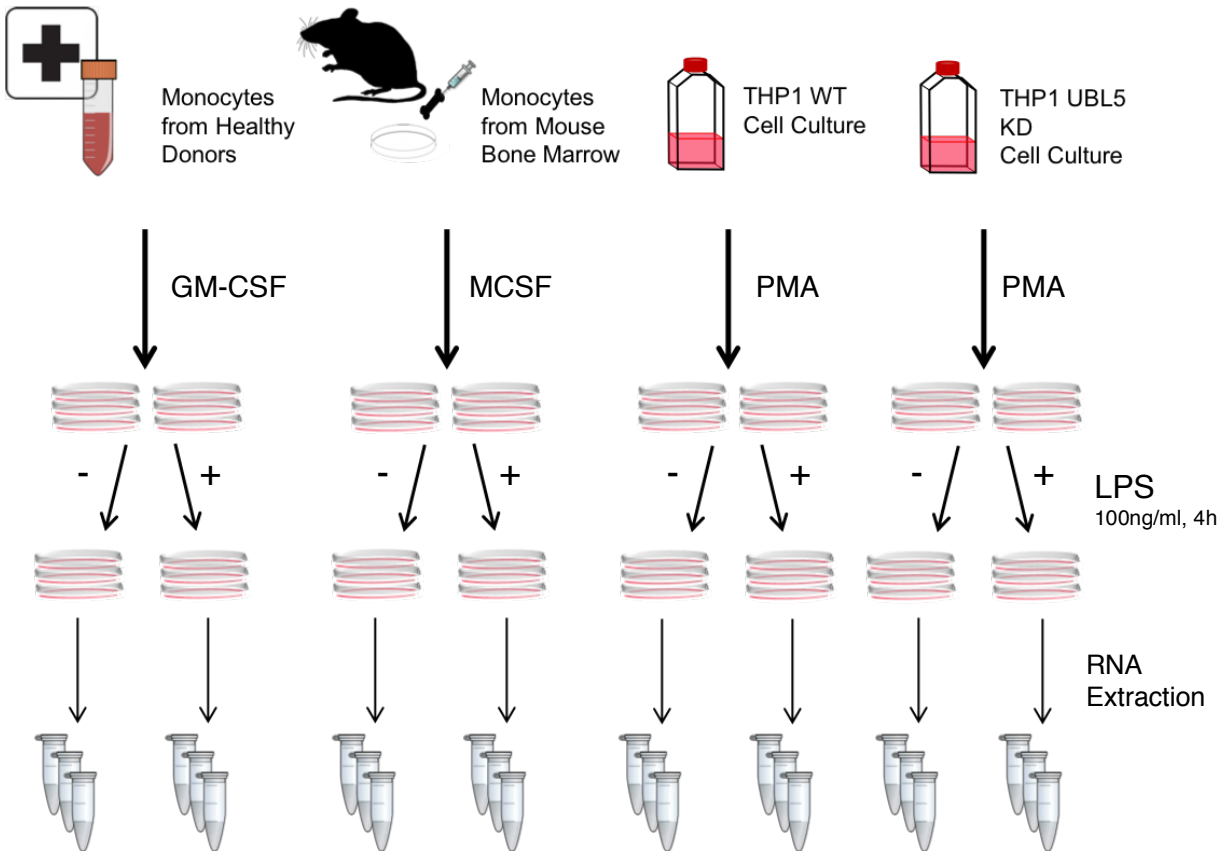


Figure 7.1: **RNA extraction from 4 macrophage cell types treated with LPS.**

Cells were collected from four sources. (L-R) Human peripheral blood monocytes collected from three screened healthy donors, bone marrow derived monocytes from C57BL/6 mice, THP-1 cells, and THP1 cells with UBL5 knockdown by siRNA. Cells were differentiated to macrophages in culture. Following differentiation, cells were split into LPS treated and untreated subsets. Following four hours of LPS treatment all cells were lysed and the RNA extracted. Three replicates of each condition and cell type was collected.

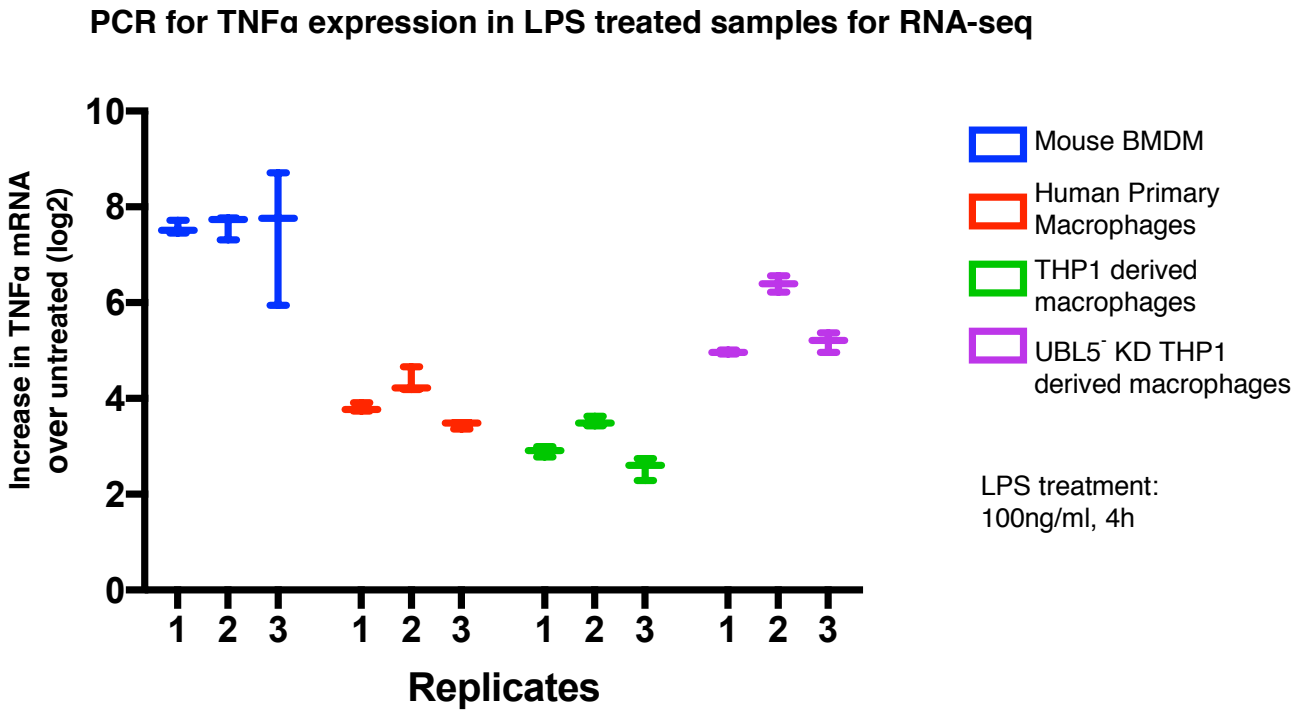


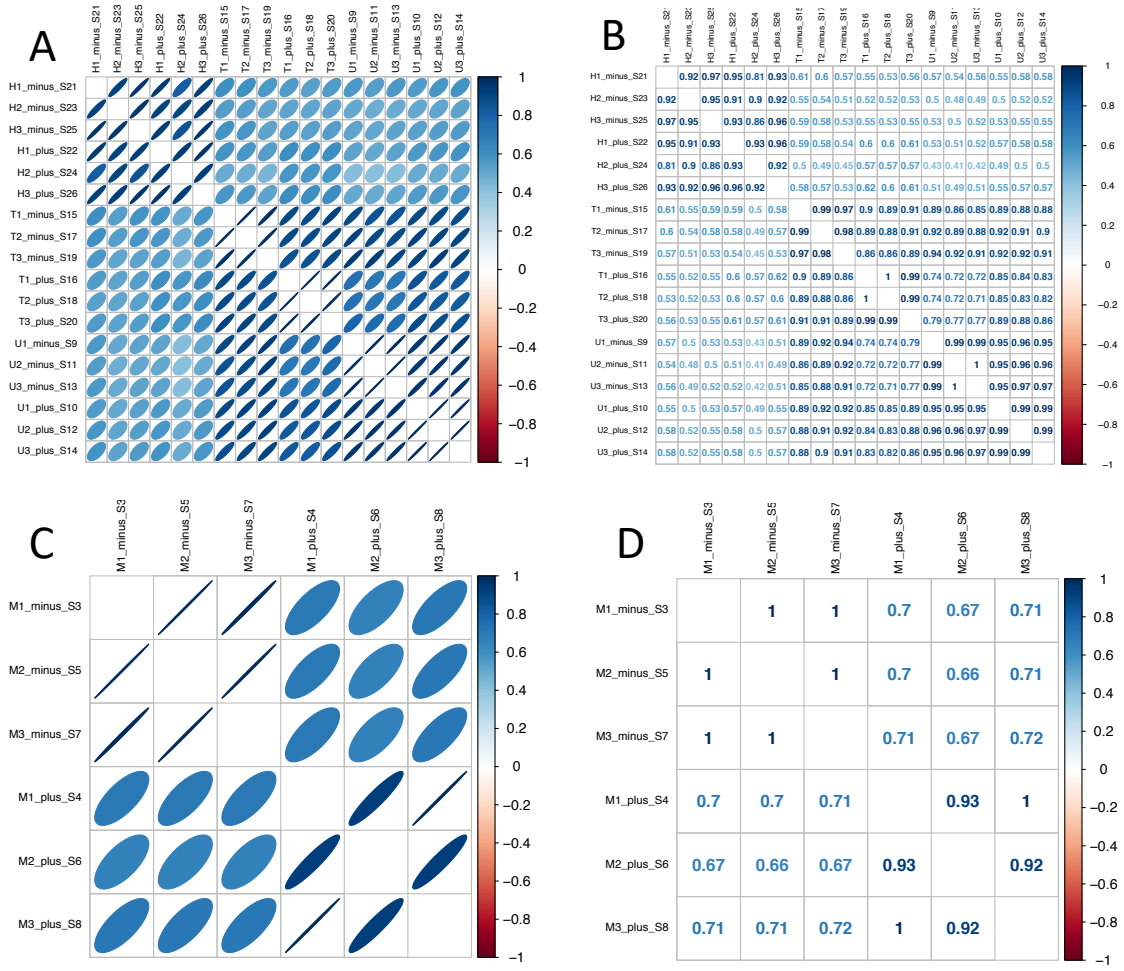
Figure 7.2: **PCR of TNF transcription in LPS treated samples selected for RNA-seq.**

PCR reaction of the three replicates from each cell type treated with LPS and selected for RNA-seq. The LPS treated cells (100ng/ml, 4h) are compared to the LPS untreated cells. The change in TNF-α expression between the treated and untreated samples were measured and normalized to the expression of HPRT in treated and untreated samples. Each biological replicate was run as three technical replicates. Error bars represent the standard deviation of the three technical replicates.

7.3 LPS induces differential gene expression within and beyond the canonical TLR4 Pathway

To first observe how the replicates and different samples compare to each other, together with my collaborators at CCR-SF and the NCBR, we did a gene expression concordance analysis. Calculating the level of overlap in gene expression across the different replicates and samples showed strong agreement across replicates and critical divergence between treated and untreated cells (Figure 7.3A-D). Of note, the replicates of mouse BMDMs showed the highest level of overlap as compared to intra-condition replicates of the other macrophage cell types. While this is to be expected (BMDMs were collected from same age mice of the same genetic background, housed in the same facility, and collected on the same day), it portends that many of the fold change differences that can be measured are more likely to cross thresholds of significance in the BMDM samples as compared to samples with less comparable replicates (Figure 7.3C-D).

Gene expression concordance across macrophage samples and replicates



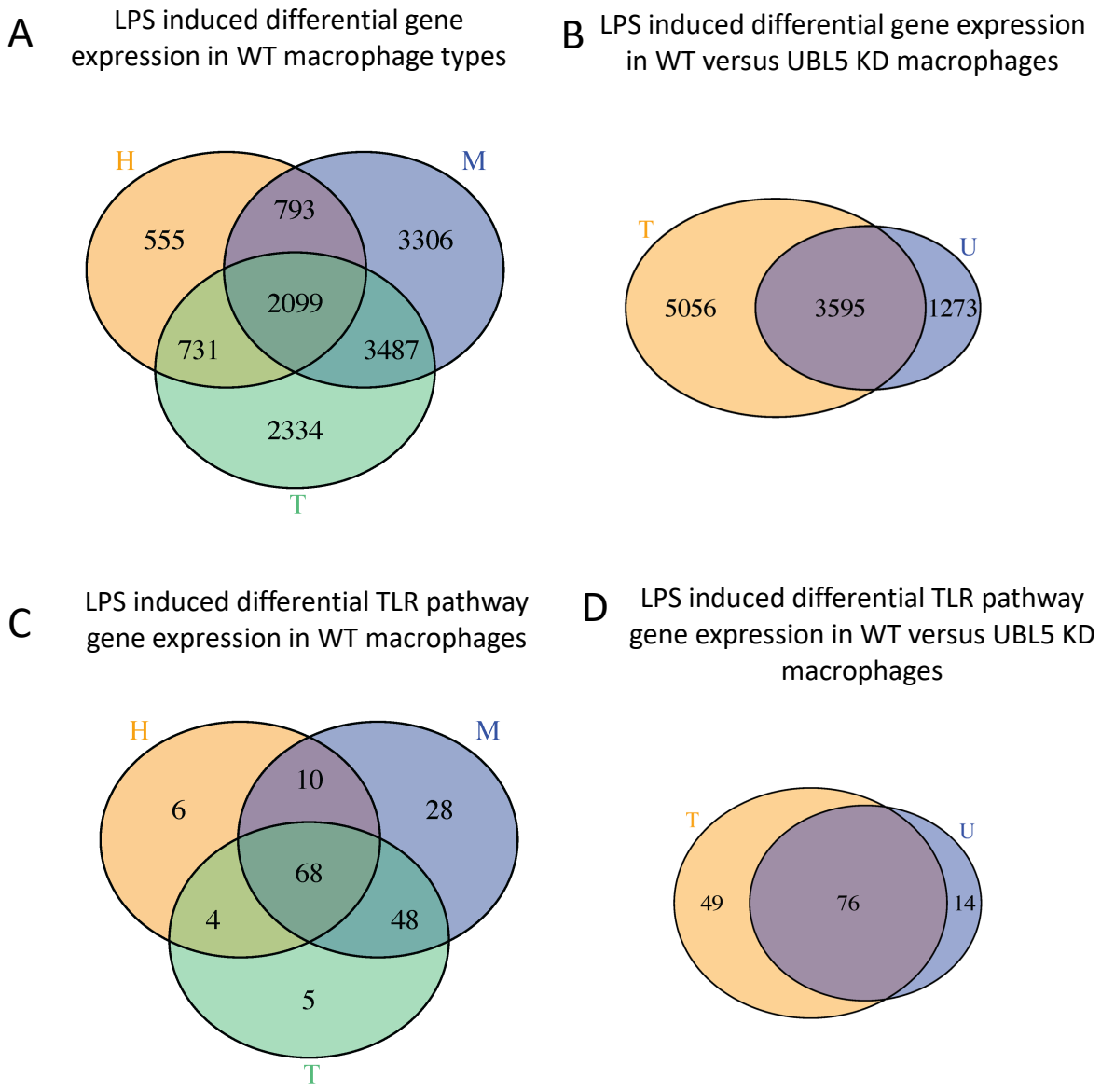
H = Human Primary Macrophages. **T** = THP1 derived macrophages.
U UBL5^{-} KD THP1 derived macrophages. **M** = Mouse BMDM.

Figure 7.3: **Concordance of gene expression profiles as measured by RNA-seq.**

Overlap of raw counts of gene expression by RNA-seq. LPS treated cells are denotated as ”plus” and untreated cells as “minus”. (A-B) Human primary cells derived macrophages (H), THP1 macrophages (T), and UBL5^{-} KD THP1 macrophages (U) are compared. Visual representation of concordance and calculated values are in the left and right panel, respectively. (C-D) Concordance of gene expression in mouse BMDMs (M) samples and replicates. Visual representation of concordance and calculated values are in the left and right panel, respectively.

Mapping the differentially expressed genes in untreated versus LPS treated cells shows that, in all macrophage samples, LPS induces differential gene expression in a sizable selection of the genome. More than 2000 of the genes with differentially expression in response to LPS are shared across all wildtype macrophage types (Figure 7.4A). Comparison between the wildtype THP1 cells and UBL5 knockdown THP1 cells also shows that more than half of the differentially expressed genes in THP1 macrophages require UBL5 to support the differential expression (figure 7.4B). This provides further support for the importance of dynamic splicing to promote and sustain the LPS-induced gene expression program. Comparison of differentially expressed canonical TLR pathway genes shows a similar trend. Human primary macrophages, BMDMs, and THP1 macrophages show differential gene expression in many of the canonical TLR pathway genes, with 68 out of 186 genes changed in macrophages from human, mouse, and immortalized human monocytes (Figure 7.4C). Similar to what we saw with global gene expression, when comparing wildtype THP1 macrophages to UBL5 knockdown macrophages a substantial portion of canonical TLR genes did not show LPS induced differential gene expression in the absence of UBL5 expression (Figure 7.4D). Since LPS induces gene expression alteration for such a large proportion of the signaling components of the pathway, this already implies that the 'state' of the TLR4 pathway in naïve cells is substantially altered after macrophage activation. It is tempting to surmise that an important theme of this change in state could involve the expression of different splice variants that have altered signaling properties that serve to maintain the post-LPS state of the cells.

Divergence and scale of LPS induced differential expression in macrophages of diverse origin.



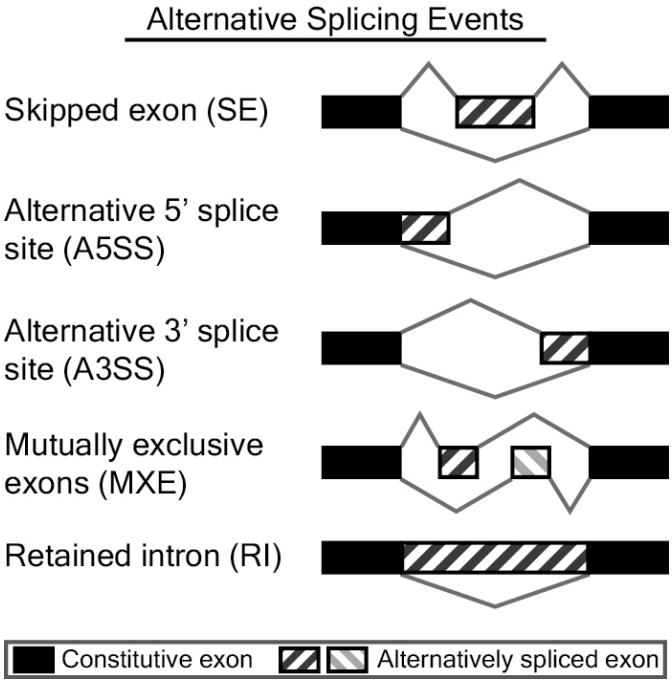
H = Human Primary Macrophages. **T** = THP1 derived macrophages.
U UBL5^{-} KD THP1 derived macrophages. **M** = Mouse BMDM.

Figure 7.4: **Differential gene expression in LPS treated macrophages.**

(A) Scale of shared and divergent counts of differentially expressed genes in LPS treated human primary cells derived macrophages (H), Mouse BMDMs (M), and THP1 macrophages (T). (B) Scale and overlap of differentially expressed genes in LPS treated wildtype THP1 macrophages (T) and UBL5^{-} KD THP1 macrophages (U). (C-D) Canonical TLR pathway genes that have differential expression in response to LPS treatment.

7.4 Diversity of splicing events are consistent across species and cell type

To investigate the alternative splicing events induced by LPS I worked with Dr. Justin Lack and Dr. Cihan Oguz from the NCBR group to apply the replicate multivariate analysis of transcript splicing (rMATS) analysis pipeline to our data. The rMATS analysis pipeline uses the comparison of two sequencing samples to map the type of alternative splicing events that explain the transcript differences between them (Shen et al., 2014). rMATS curates splicing events into five alternative splicing event types. Skipped exon (SE) which involves the skipping of an exon between two transcripts. Alternative 5' splice site (A5SS) which involves variation in the 5' region by the preference for alternative splice sites near the 5' site. Alternative 3' splice site (A3SS) which involves variation in the 3' region by the preference for alternative splice sites near the 3' site. Mutually exclusive exons (MXE) which occurs when two exons are interchangeably included in transcripts. Retained introns (RI) which is when an intronic sequence is included by alternative splicing in a transcript. (Figure 7.5).



From Park, J. W., Tokheim, C., Shen, S., & Xing, Y. (2013).

Figure 7.5: **Classification of alternative splicing event types.**

Alternative splicing event types and their abbreviations as defined by the rMATS analysis platform. From top to bottom) Skipped exon (SE), Alternative 5' splice site (A5SS), Alternative 3' splice site (A3SS), Mutually exclusive exons (MXE), and Retained intron (RI).

A cursory analysis of splicing events induced by LPS treatment in the four samples I measured found that all five splicing events occurred in near similar proportions in the differently sourced macrophages. Splicing events were dominated by skipped exon events with smaller proportions of the other four event types. These ratios were, interestingly, not effected by the knockdown of the splicing factor UBL5 (Figure 7.6). The most observable difference between the different macrophage samples were the number of splicing events in response to LPS treatment. BMDMs and THP1 macrophages showed the greatest number of alternative splicing events in response to LPS treatment. Human primary macrophages and UBL5 knockdown THP1 macrophages show only about half as many splicing events in response to LPS treatment as THP1 wildtype and BMDMs (Figure 7.6). Looking at the genes associated with those splicing events reveals that a substantial number of the genes undergoing alternative splicing are shared across samples (Figure 7.7A). Across the wildtype sample (Human primary, BMDMs, THP1) 32 genes with alternative splicing events are shared (Figure 7.7B). A list of the 32 genes reveals several TLR pathway related genes, such as IRAK1, TANK, and CASP8. The regulatory factor MBNL1 that was found to be critical for macrophage differentiation and regulating splicing events (H. Liu et al., 2018) is also found on the list as undergoing alternative splicing in response to LPS (Figure 7.7C). The mitochondrially targeted NDUFAF6 is also listed amongst the genes undergoing alternative splicing in response to LPS suggesting a mitochondrial and metabolic regulatory link. (Alternatively spliced variants of this gene have also been associated with a form of Fanconi Syndrome a chronic kidney disease associated with pulmonary interstitial fibrosis (Hartmannová et al., 2016)).

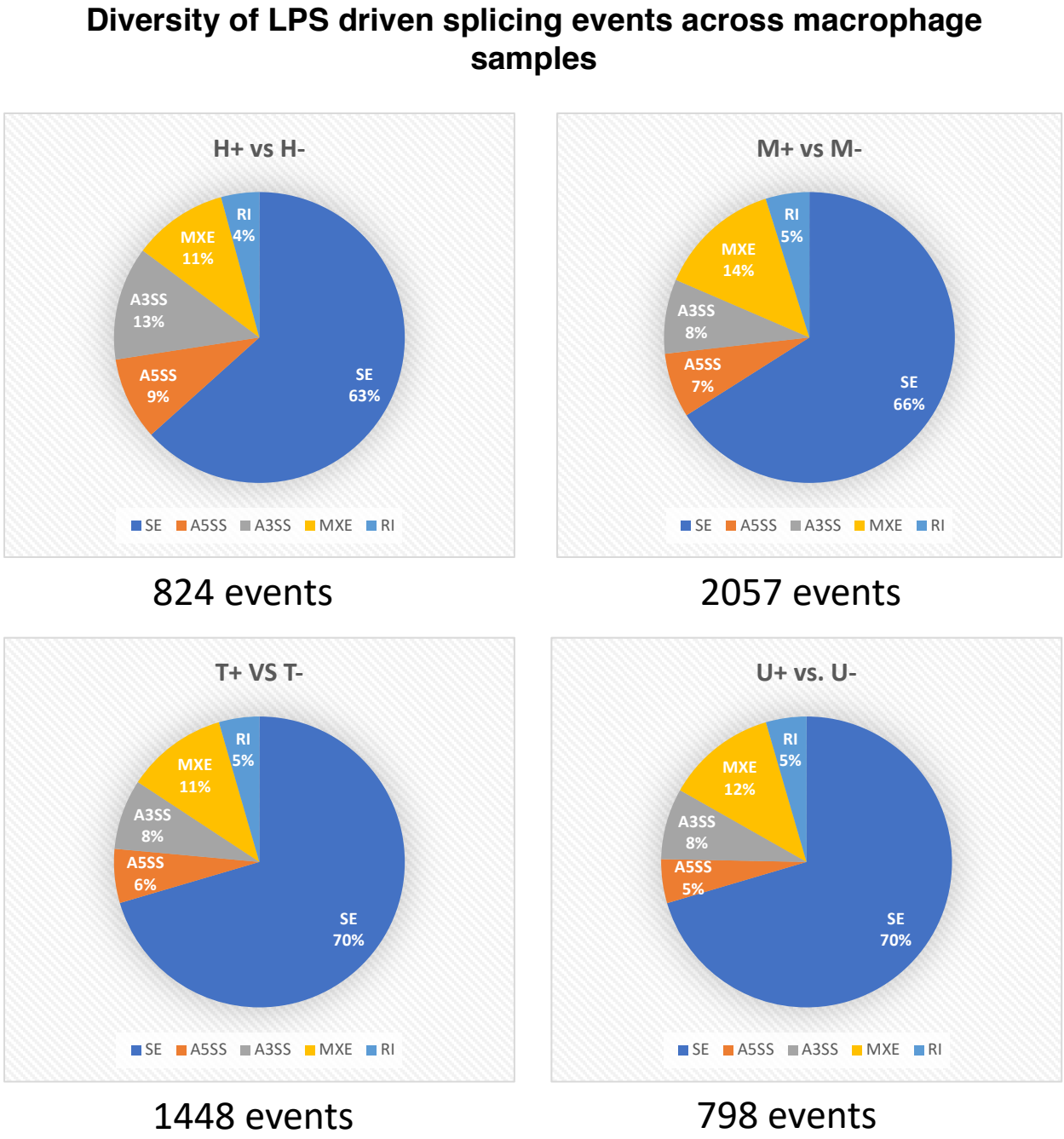
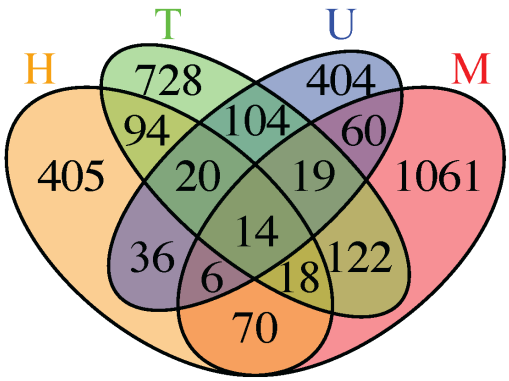


Figure 7.6: **Splicing events in different macrophage cell types treated with LPS.**

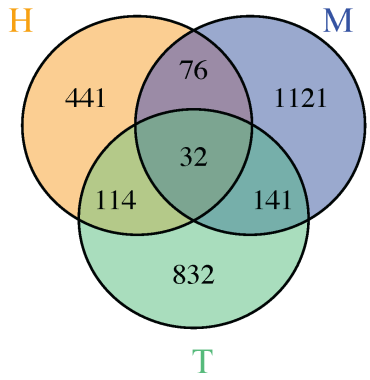
Characterization of splicing events as found in rMATS analysis of RNA-seq data from human primary cells derived macrophages (H), Mouse BMDMs (M), THP1 macrophages (T), and UBL5 KD THP1 macrophages (U). Proportion of each event type represented in the Venn diagram and number of total events observed listed below.

Genes with alternatively spliced mRNA in response to LPS

A Number of genes with alternatively spliced mRNA



B Number of genes with alternatively spliced mRNA (Wildtype cell types only)



C Genes with alternatively spliced mRNA in all cell types (WT)

Gene	Score	Gene	Score
CD44	5.441	RAD52	0.886
LRRFIP2	3.13	RCOR3	0.857
IL15	3.028	CASP8	0.782
PGAP2	2.334	HPS5	0.776
TET2	1.563	ARRB2	0.763
NFKBIZ	1.315	CMC2	0.7
TMEM168	1.207	NDUFAF6	0.665
TTLL3	1.168	FRYL	0.594
SLC7A7	1.058	CENPT	0.586
IRAK1	1.044	MACF1	0.527
MBNL1	1.005	FAM49B	0.527
ENOX2	0.987	ZNF195	0.5
ACAD10	0.948	BRAF	0.454
ST7L	0.939	TTC13	0.426
DPEP2	0.916	AASDH	0.424
TANK	0.9	AP1AR	0.22

Figure 7.7: **Alternative splicing events in shared genes following treatment with LPS.**

(A) Number of genes with observed splicing events in response to LPS treatment found in human primary cells derived macrophages (H), Mouse BMDMs (M), THP1 macrophages (T), and UBL5^{-} KD THP1 macrophages (U). (B) Number of genes with observed splicing events in response to LPS treatment found in the wildtype cell line derived macrophages. (C) Table of shared genes with measured splicing events. Score is calculated by the splicing event with the highest "inclusion level difference".

7.5 In human and mouse macrophages, genes related to LPS signaling are alternatively spliced in response to LPS

The extent of conservation between human and mouse TLR4 signaling architecture has been the subject of much study and debate (Godec et al., 2016; Seok et al., 2013; J. Sun et al., 2016; Takao & Miyakawa, 2015) (Figure 1.3). While some core signaling factors are conserved across species, other factors have been shown to be species specific (such as IRAK1 and IRAK2, (J. Sun et al., 2016)). Alternative splicing has also been shown to operate in a species specific manner, with conserved genes generating different isoform variations across species (Pan et al., 2005). In the previous section I have shown that in response to LPS many genes are alternativly spliced in macrophages from primary human and mouse monocytes. (The higher proportion of genes with observed alternative splicing events in mouse versus human may in part be driven by the closer similarity of its replicates. As discussed in section 7.3, Figure 7.3). Genes undergoing alternative splicing in response to LPS only modestly overlap across human and mouse (Figure 7.8A). An analysis of pathway enrichment, however, shows that in both species the genes undergoing alternative splicing in response to LPS are in pathways related to TLR signaling (Figure 7.8B). In human, of the 70 pathways identified as enriched for in the alternatively spliced genes, 34 are in TLR related pathways. In mouse, of the 136 pathways enriched for in the alternatively spliced genes, 52 are in pathways associated with TLR signaling. Critically, the 12 pathways that are enriched for in both species are dominated by TLR related signaling pathways (Figure 7.8B). These findings suggest that the engagement of alternative splicing in response to LPS is conserved across species, and that its activation effects a broad range of TLR signaling genes.

LPS induced alternative splicing in human and mouse macrophages

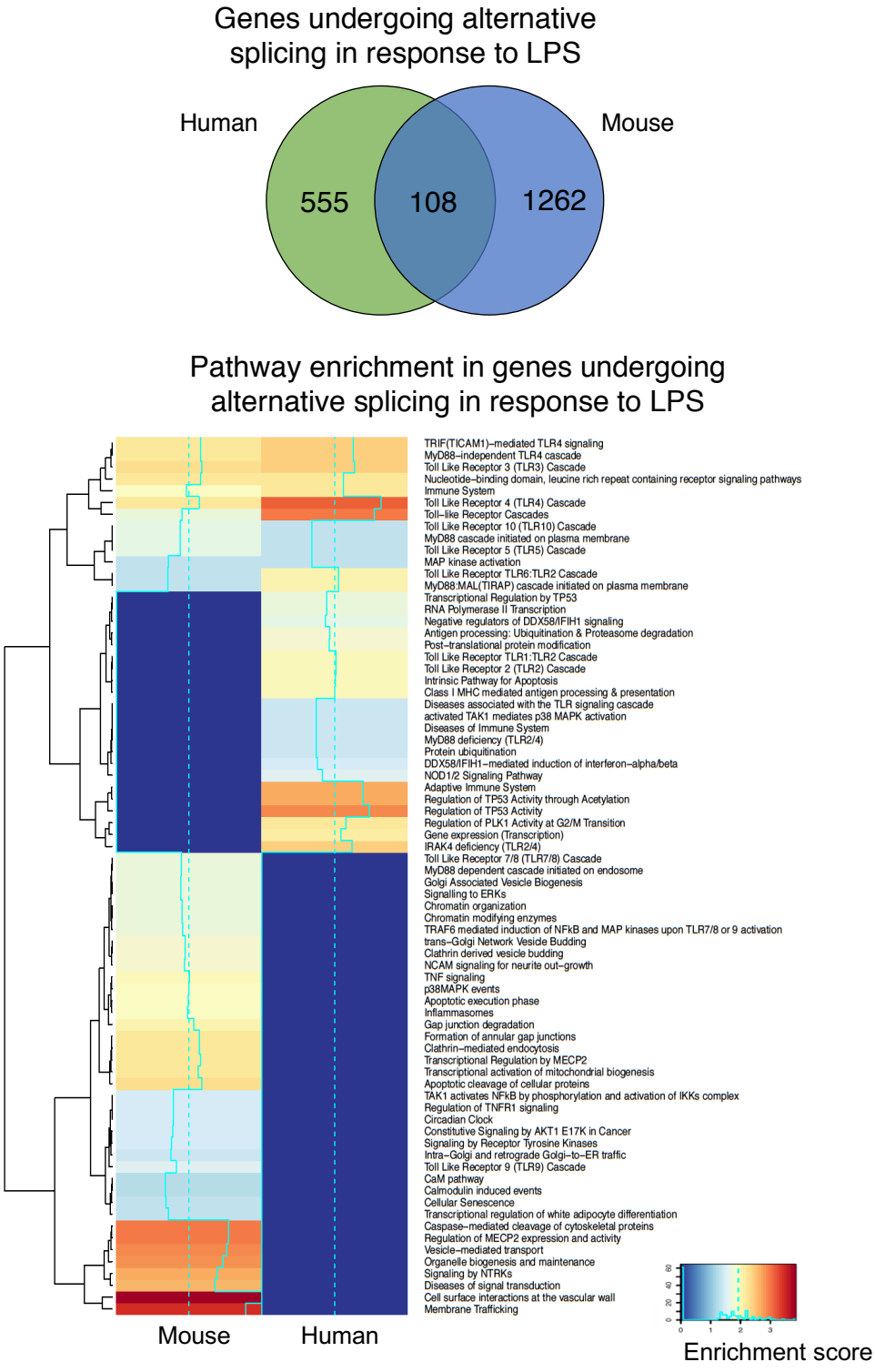


Figure 7.8: **Shared enrichments in human and mouse macrophages of LPS induced alternative splicing.**

(A) Number of genes with observed splicing events in response to LPS treatment found in human primary cells derived macrophages and Mouse BMDMs. (B) Reactome pathway enrichments (p<0.05) of genes in human and mouse macrophages that undergo alternative splicing events in response to LPS.

7.6 Long read sequencing identifies LPS driven differences in repertoire of mRNA transcripts

The power of short read bulk RNA-seq is that it provides a quantitative measure for the mapped splicing events it detects. This approach, however, relies on the short reads (usually of about 150 base pairs (bp)) being mapped onto a known reference genome so that the expressed transcripts can be reconstructed. The dependency on the reference genome means that short read RNA-seq is hindered from identifying novel transcripts that are as yet not curated in the reference genome and its variants. An alternative approach is long read RNA-seq methods that can read upwards of 5000bp at a time, thus requiring no reconstruction of splice junction to infer transcript variants. A table of the differences between short and long read approaches are listed in table 7.1. To complement the short read RNA-seq analysis, I collaborated with the CCR-SF group to conduct long read RNA-seq by SMRT PacBio Sequencing (Table 7.2). A representative single replicate RNA sample set from the short read analysis was selected for long read sequencing. To compare the various macrophage samples, in the context of novel isoforms, PacBio results were analyzed by the Structural and Quality Annotation of Novel Transcript Isoforms (SQANTI) pipeline (Tardaguila et al., 2018). The SQANTI pipeline still uses the reference genome, not as a way to map and reconstruct transcripts, but to categorize how the identified transcripts relate to known transcripts. SQANTI categorizes transcripts as full splice match (FSM), incomplete splice match (ISM), novel in catalog (NIC), novel not in catalog (NNC), genic intron, and genic genomic (Figure 7.9). A SQANTI analysis of the treated and untreated macrophages I selected showed high levels of FSM and ISM transcripts, with lower levels of the other categories. Of interest, the number of incomplete splice matches was substantially higher in the treated and untreated UBL5 knockdown samples, further suggesting that a segment of the alternatively spliced variants we observe in macrophages depend on UBL5 (Figure 7.10).

	Illumina short read sequencing	SMRT Pacific Biosciences (PacBio)
Base pairs/read	~150	>5,000
Novel isoforms not in the reference genome	Relies on matching to reference genome	Can identify novel isoforms
Scale	Affordable to scale to multiple conditions and replicates	Expensive to expand to multiple conditions and replicates

Table 7.1 **Short read versus long read RNA-seq methods.**

	Illumina short read sequencing	SMRT Pacific Biosciences (PacBio)
Cell types	H, M, T, U	H, M, T, U
Conditions	LPS-/LPS+	LPS-/LPS+
Replicates	3	1

Table 7.2 **Sample and replicate conditions for short read and long read RNA-seq.**

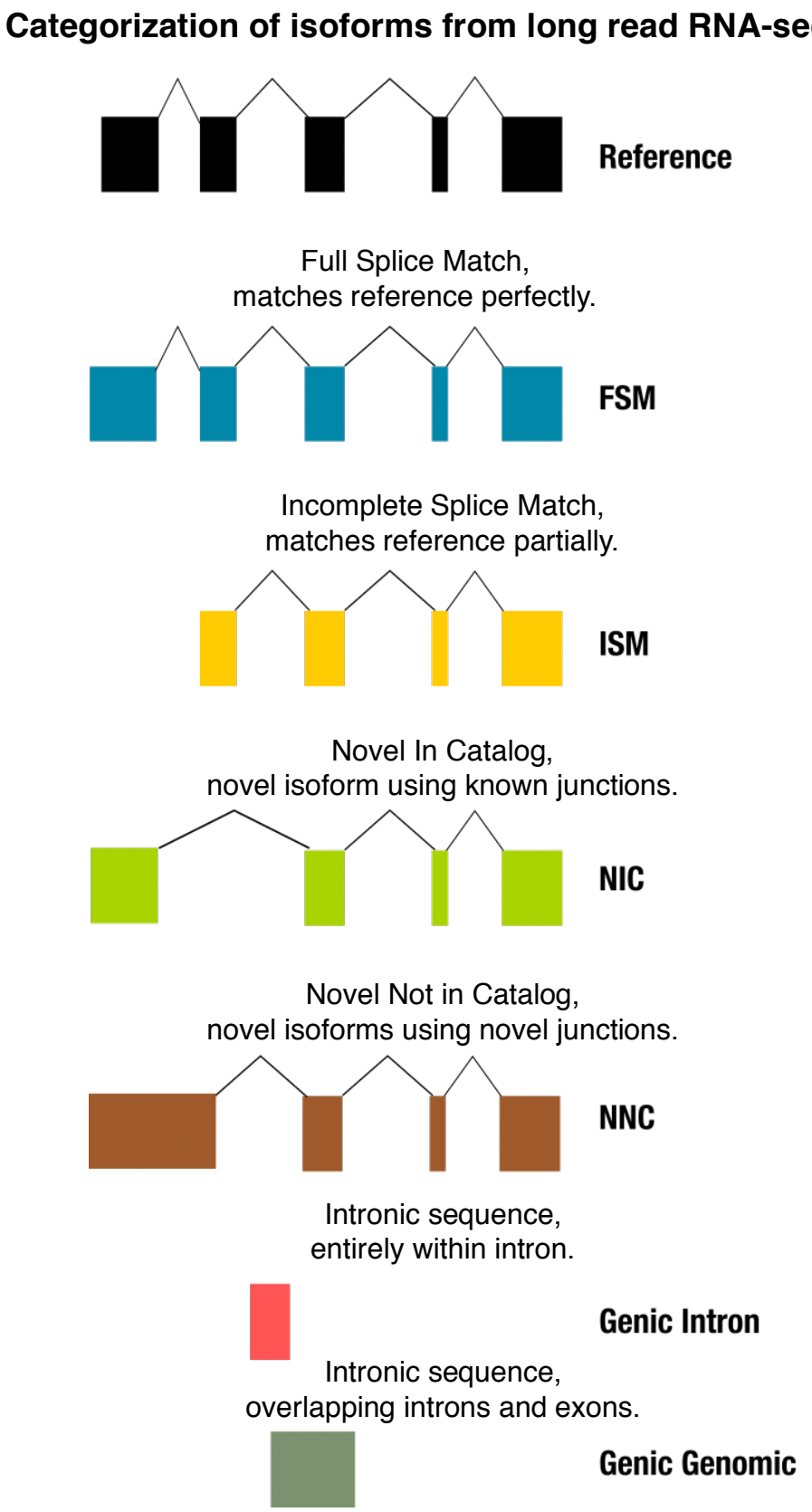
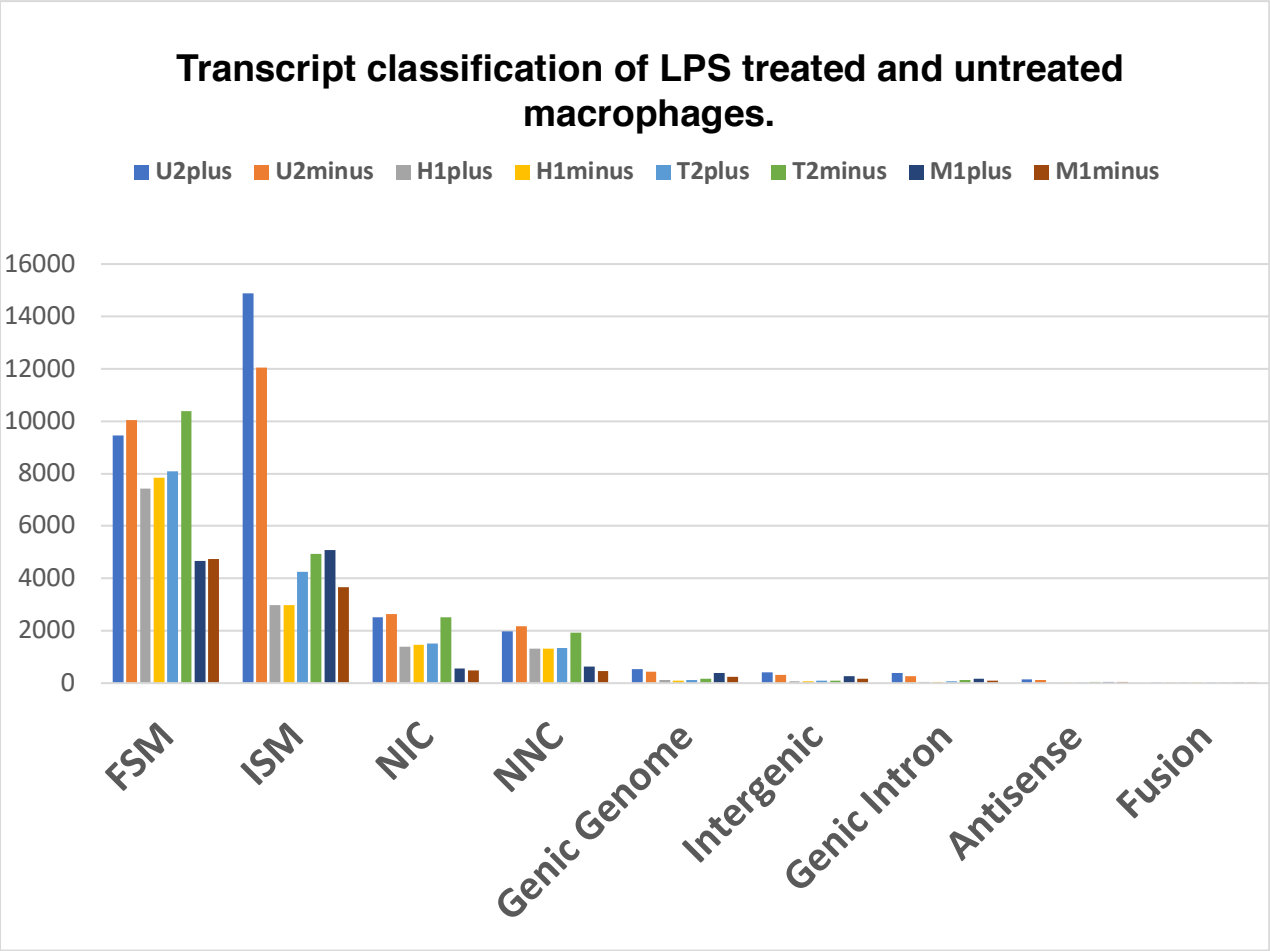


Figure 7.9: **Novel isoform categorization by SQANTI analysis of long read RNA-seq.**

A schematic of the categorization of isoforms from long read RNA-seq. Visual mapping is of how the different isoforms map onto the reference isoform (black). Image adapted from *https://www.pacb.com/wp-content/uploads/2018-SMRT-Leiden-Iso-Seq-Method-Human-Diseases-Elizabeth-Tseng-MainTalk.pdf*



H = Human Primary Macrophages. **T** = THP1 derived macrophages.
U UBL5^{-} KD THP1 derived macrophages. **M** = Mouse BMDM.

Figure 7.10: **Classification of transcripts identified by long read RNA-seq of LPS treated and untreated macrophages.**

LPS treated samples are indicated as “plus” and untreated as “minus”. Numbers reflect total different transcripts identified. Categories follow SQANTI categorization (see figure 7.8)

7.7 Summary

The identification of the spliceosome as a regulatory signaling mechanism in the context of macrophage activation implies a dynamic macrophage transcriptome in the early stages of the LPS response. To elucidate how these changes in mRNA transcripts activate and sustain LPS signaling requires a robust characterization of the alternative splicing events and novel transcripts induced by TLR4 signaling. In this chapter I have demonstrated the creation of such a dataset by creating complimentary short read and long read RNA-seq analysis. The diversity of macrophage samples used in the dataset further enables cross-species and cell line comparison for identifying variant changes that are relevant in human disease. Critical interrogation of these datasets, paired with rigorous follow up studies, is still required. These datasets outline a path towards further characterizing the novel regulatory mechanism I have shown in chapter 6.

8 Discussion

8.1 Outline

Taken together, the findings that I have presented in the preceding chapters suggest two parallel sets of insights. First, a framework for analysis of genome-scale data that goes further than previous methods in identifying lower scoring true positive hits. Second, a regulatory role for the spliceosome and proteasome in initiating and sustaining macrophage activation on a more detailed scale than has been previously appreciated. In this chapter I discuss some of the implications of these two findings. I begin with a summary of the different findings of this thesis (section 8.2). I discuss some of the questions and possible follow up studies that the two major findings suggest (section 8.3 and section 8.4). Based on the latter finding from this thesis I also propose a new model for how stimulated macrophages sustain TLR4 signaling against a background of active regulatory expression (section 8.5). As proposed by the model, a role for alternative splicing in the regulation of inflammatory signaling would elucidate some of the innate immune signatures that are observed in diseases associated with splicing-related mutations and differential isoform expression. In the final section I present one such example and how the work of this thesis can be built upon to expand our understanding of the mechanism behind a specific disease pathology (section 8.6).

8.2 Summary of the work presented in this thesis.

The work I have presented in this thesis support three hypotheses about high-throughput data analysis and the regulation of TLR4 signaling. First, that pathway and network approaches to hit selection are complimentary in the solutions they provide, and a systemic integration of these approaches would yield cumulative results. Second, that a robust analysis of genome-scale studies that goes beyond the highest scoring and recognizable hits would reveal novel

regulatory mechanisms of macrophage activation. Third, that the spliceosome and alternative splicing play a dynamic and potentially critical role in macrophage activation.

Using previously published datasets for HIV HDFs, I first developed and tested a novel integrated and iterative framework, coined TRIAGE, for using curated databases in hit selection from genome-scale screens. By measuring the statistical significance of enrichment and the size of overlap across parallel omic studies, I showed that pathway and network-based approaches are indeed complimentary with strengths in false positive and false negative correction respectively (Sections 4.1-4.9). I then showed how the TRIAGE framework for integrating pathway and network-based hit selection to prioritize hits from a two cutoff dataset leads to the strongest combined error correction (Sections 4.10-4.13).

Applying the TRIAGE approach to hit selection from three genome-scale siRNA studies of the response to LPS in macrophages I identified many shared putative regulators with a strongly improved enrichment of TLR and other expected innate immune pathways (section 6.3). The analysis also identified a robust enrichment of splicing and proteasome factors (section 6.4). The integrated approach further identified putative regulatory and interacting candidates of these processes that are currently outside of the annotated spliceosome and proteasome groups (section 6.7). Using a proteasome inhibitor, I showed how the proteasome is more broadly required in the downstream signaling branches of the TLR pathway than has previously been appreciated (section 6.5). Using a splicing inhibitor, I also showed how the activation of the TNF-α promoter by LPS requires an active spliceosome (section 6.6). I further demonstrated that TLR factors with known variants show differential exon usage in response to LPS stimulation (section 6.8). To expand the capacity for characterizing the role of alternative splicing in macrophage activation I generated and mapped a set of RNA-seq datasets from human and mouse macrophages treated with LPS (section 7.2). Initial global analysis of alternative splicing and the increase in splicing events in response to LPS

complemented my earlier finding of the likely important role of the spliceosome in macrophage activation, especially considering the high frequency of TLR pathway components among the spliced transcripts (sections 7.4-7.5). Long read RNA-seq broadens this dataset by identifying novel transcripts not currently annotated in the reference genome (section 7.6). These two datasets suggest a pathway through which the quantitative data from the short-read RNA-seq can be integrated with the robust transcript detection of the long-read approach to achieve a more comprehensive characterization of the changes that occur in the macrophage transcriptome in response to LPS.

In the processes of developing the above analysis, I also recognized that the TRIAGE pipeline has a utility for the high-throughput biology research community beyond the innate immune context for which I developed it. To broaden its accessibility, I built a publicly available web-based interface of the analysis pipeline (chapter 5). A version of the interface was first made publicly available in February 2018. Since then, *triage.niaid.nih.gov* has had more than 1,000 unique visitors, with the majority returning for repeat visits. The most common question raised by users has been the challenge in how to tier the high confidence and medium confidence datasets in different contexts. The ease of use and speed of analysis by TRIAGE is designed to make it possible to address the above issue through repeat analysis with different cutoffs to compare and contrast different analysis outcomes (section 5.11). The findings of this thesis also raise critical questions and suggest paths for further exploration both in bioinformatic solutions to high-throughput guided discovery and in determining the role of alternative splicing in the context of inflammatory signaling and disease. In this chapter I address some of the possible paths that follow from this analysis.

8.3 Persistent challenges in database-driven omics exploration

In the development of TRIAGE, I have focused on the twin challenges of incomplete genome annotation by pathway databases and lack of false positive correction in network analysis approaches. TRIAGE addresses these issues and optimizes the use of databases to more thoroughly identify biologically significant lower scoring hits from omic-scale studies. TRIAGE was designed and tested using the first-generation methods of pathway and network analysis (over representation analysis and direct neighbor, respectively). While the framework of the complementary iterative design of TRIAGE can in principal be extended to other forms of pathway and network analysis, it remains to be tested whether the design can be similarly applied to those analysis methods.

Some of the intrinsic challenges of relying on curated databases persist even in the TRIAGE design. In the context of using pathway enrichment for hits prioritization, the statistical approach used to assess significant enrichment (a hypergeometric test with FDR) favors a specific range of pathway sizes. Best practices in the use of pathway enrichment statistics suggest that the analysis works best in pathways that contain member genes in the range of 20 to 400 (Ramanan et al., 2012). This range limits the possibility to explore broader pathways such as metabolic processes which have higher gene membership counts. There is also a lot of redundancy and overlap in pathway annotation which can lead to unrelated enrichments being identified, though some bioinformatic solutions for this have already been proposed (Pita-Juárez et al., 2018; Vivar et al., 2013). Network analysis driven data exploration also has a set of persistent challenges. Notably, network analysis databases such as STRING are cell type and treatment agnostic (Ma et al., 2019), making some of the imputed interactions irrelevant or misleading for analysis in different contexts. The latter challenge could be addressed by more cell or disease specific protein-protein interaction networks being generated, but this will require a substantial investment from the research community to

develop and generate such resources. I designed the TRIAGE R code such that it can be adapted to more bespoke analysis pipelines when such datasets are available. Recently developed search engines such as GADGET have used thoroughly sensitive text mining algorithms to map abstracts in PubMed to specific gene IDs, metabolites, and disease keywords (Craven, 2015), an expansion of these methods to map interactions from the literature to the cell types and treatments they were identified in could address some of the current network database blindspots. Additional creative bioinformatics, however, will also be necessary to infer across which cell types and conditions observed protein-protein interactions can be extrapolated to, and in what contexts comparisons are less likely to be informative.

8.4 Alternative splicing as a regulatory mechanism for macrophage activation

The short and long read RNA-seq datasets I described in chapter 7 provide an opportunity to explore many critical questions of how alternative splicing might actively regulate macrophages following stimulation by LPS. The comparative data in human and mouse primary macrophages make it possible as a first step to first characterize how many of the genes that undergo alternative splicing in response to LPS are shared across human and mouse. Of those shared genes, an essential follow up question that these datasets can reveal is if the alternative splicing sites and mechanisms are shared or different between species. This will shed light if gene expression conserved in mouse and human TLR4 response also correlates with shared isoform expression. Another essential path of inquiry is to characterize and validate by PCR the novel isoforms and the differential alternative splicing identified by both RNA-seq methods. Commercially available antibodies are often not effective at differentiating between different isoforms (as a recent study exploring the different isoforms of TMEM173, the gene transcribing the STING protein, found (Rodríguez-García et al., 2018)), thus validation by qPCR may be the prudent method to validate the candidates from the RNA-seq findings.

Targeting the 32 shared candidates identified as alternatively spliced in wildtype macrophages (Figure 7.7) by siRNA knockdown could serve as an initial test for a regulatory role in the LPS response. Exogenous expression of specific variants in the knockdown cell lines could be further used to delineate the comparative roles of different splice variants.

The prevalence of alternatively spliced signaling factors raises the question of what the mechanistic effects of differently spliced variants imply. The diversity of alternative splicing events identified in the datasets (Figure 7.6) could serve as a starting guide for exploratory analysis. A number of recent papers, for example, have suggested that intron retention by alternative splicing is a cellular mechanism of inducing nonsense mediate decay of a protein and thereby decreasing its abundance (Ge & Porse, 2014; Jacob & Smith, 2017). Looking at this mechanism from a different angle raises questions of how contextual alternative splicing events are guided and selected for. Studies have shown that splice sites are often selected based on relative GC nucleotide richness, and GC vs AT richness is reflective of a "splice site strength" (Amit et al., 2012; M. Wang & Marín, 2006). A computational tool to asses predicted splice strength has been developed (Yeo & Burge, 2004).

The TRIAGE analysis of the three LPS screens also provides an additional approach through which to elucidate the alternative splicing regulatory mechanism. Analysis of the three LPS screens identified a network of splicing and spliceosome supporting factors, many of which function as RNA-binding proteins (RBPs) (section 6.7). Computational methods have been developed to predict binding between specific sequences and RBPs (Paz et al., 2014). These methods could be used to prioritize candidates from the LPS siRNA screen based on their predictive binding with alternatively spliced isoforms from the LPS RNA-seq studies. Identified regulatory candidates could then be experimentally tested by the use of siRNA knockdown in THP1 cells and measuring its impact on the abundance of the targeted

alternatively spliced isoform. These parallel approaches are a path by which the effect and mechanism of alternative splicing on TLR4 signaling can be systemically mapped.

8.5 A model for context specific sensitivity in macrophage activation

The critical dependency of TLR signaling activation on an active spliceosome and proteasome, as I have shown in chapter 6, suggests a possible model by which macrophages switch to sustained inflammatory activation from its homeostatic state of robust transcription of factors that suppress activation. Alternative splicing of inhibitory factors in a way that selectively targets them for degradation by the proteasome can sustain inflammatory activation even in the context of robust inhibitory transcription. Alternative splicing can also blunt the inhibitory effects of negative regulators of TLR by shifting its isoform selection to non-functioning variants. The explanatory potential of this model for macrophage activation dynamics is that it provides a novel avenue through which activation sensitivity, response duration, and rapid resolution can be tuned in a context specific manner. Both proteasome activation and alternative splicing have been previously shown to confer tissue specific behavior in different contexts (Badr et al., 2016; Grosso et al., 2008; Kniepert & Groettrup, 2014; Morozov & Karpov, 2019) (also discussed in section 1.4.3). Context specificity is critical for TLR signaling sensitivity, as encounters with different levels of endotoxin in different tissue contexts signal different levels of danger. In this understanding, the TLR pathway is maintained under tonic negative regulation by the robust transcription of factors that suppress its activation. When a contextually appropriate threshold of danger is sensed, the proteasome and spliceosome alter the mRNA profile and protein abundance of pathway effectors allowing signaling to successfully proceed down the pathway uninhibited.

8.6 A mechanism for inflammatory dysregulation in myelodysplastic syndromes

Myelodysplastic syndromes (MDS) is a premalignant hematological disease that is associated with chronic TLR signaling. As I review in section 1.4.4, patients with MDS have a high prevalence of mutations in various core splicing factors. A model of spliceosome and proteasome driven regulation of inflammatory activation would suggest a possible mechanism for the dysregulated inflammation found in patients with MDS. Two studies looking at MDS patients with spliceosome mutations found that patient cells expressed different isoforms of critical TLR pathway effectors. Analysis of IRAK-4 expression found that MDS patients with mutations in the splicing factor U2AF1 express a long form of IRAK-4 (IRAK4-L) that constitutively activates NFκB (Smith et al., 2019). In the context of the findings I have presented, and the proposed model for TLR regulation, this could suggest that in the non-disease state macrophages toggle between the activating and non-activating isoforms of IRAK-4 to control the inflammatory response. In MDS patients with splicing related mutations this capacity may be compromised, leading to the dysregulated TLR signaling that is observed.

Another study of MDS patients found that hnRNPA1, an auxiliary splicing factor (discussed in section 1.4.3), leads to alternative splicing of ARHGAP1 which then activates the GTP-binding Rho family protein Cdc42 and is associated with dysregulation in HSC of MDS patients (Fang et al., 2017). This finding further shows how dysregulation of a splicing factor is correlated with aberrant innate immune signaling. Intriguingly, close observation of the targets I identified by TRIAGE analysis of genome-scale studies of the response to LPS suggests a possible mechanism that can fill in some of the steps that such a process might take. In Figure 6.7 (*p.* 151) I mapped the canonical TLR pathway and spliceosome pathways hits from the screen together with hits from the screen that have predicted interactions with the pathway members. A study of the resulting network shows that PTKB which is activated by

LPS interacts with Paxillin which also interacts with GIT1, both of which are hits identified in my analysis of the screens by TRIAGE. GIT1 interacts with another hit identified in the screen, ARHGEF6. ARHGEF6 is a reciprocal modulator of ARHGAP1 and functions as a GTPase exchange factor with Cdc42. ARHGAP1 and Cdc42 are the two factors that the study in MDS patients found to be effected by the splicing factor and associated with dysregulated HSC. This analysis suggests a pathway by which the findings I have shown in this thesis can be used to further elucidate the mechanism of inflammatory signaling in health and disease. Building on this work, I have begun a collaboration with Dr. Neil Young at the National Heart Lung and Blood Institute at the NIH to compare the RNA-seq data from macrophages derived from the cells of MDS patients with the findings I have shown in Chapter 7 of this thesis.

9 References

Aguirre, A. J., Meyers, R. M., Weir, B. A., Vazquez, F., Zhang, C.-Z., Ben-David, U., . . . Doshi, M. B. (2016). Genomic copy number dictates a gene-independent cell response to CRISPR/Cas9 targeting. *Cancer discovery, 6*(8), 914-929.

Akira, S., Takeda, K., & Kaisho, T. (2001). Toll-like receptors: critical proteins linking innate and acquired immunity. *Nature Immunology, 2*(8), 675-680. doi:10.1038/90609

Akira, S., Uematsu, S., & Takeuchi, O. (2006). Pathogen Recognition and Innate Immunity. *Cell, 124*(4), 783-801. doi:https://doi.org/10.1016/j.cell.2006.02.015

Alexander, C., & Rietschel, E. T. (2001). Invited review: bacterial lipopolysaccharides and innate immunity. *Journal of endotoxin research, 7*(3), 167-202.

Alexopoulou, L., Holt, A. C., Medzhitov, R., & Flavell, R. A. (2001). Recognition of double-stranded RNA and activation of NF-kappaB by Toll-like receptor 3. *Nature, 413*(6857), 732-738. doi:10.1038/35099560

Amit, M., Donyo, M., Hollander, D., Goren, A., Kim, E., Gelfman, S., . . . Ast, G. (2012). Differential GC Content between Exons and Introns Establishes Distinct Strategies of Splice-Site Recognition. *Cell Rep, 1*(5), 543-556. doi:https://doi.org/10.1016/j.celrep.2012.03.013

Anders, S., Reyes, A., & Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. *Genome Research, 22*(10), 2008-2017. doi:10.1101/gr.133744.111

Arakelyan, A., Nersisyan, L., Poghosyan, D., Khondkaryan, L., Hakobyan, A., Löffler-Wirth, H., . . . Binder, H. (2017). Autoimmunity and autoinflammation: A systems view on signaling pathway dysregulation profiles. *PLOS ONE, 12*(11), e0187572. doi:10.1371/journal.pone.0187572

Arango Duque, G., & Descoteaux, A. (2014). Macrophage cytokines: involvement in immunity and infectious diseases. *Front Immunol, 5*, 491. doi:10.3389/fimmu.2014.00491

Arber, D. A., Orazi, A., Hasserjian, R., Thiele, J., Borowitz, M. J., Le Beau, M. M., . . . Vardiman, J. W. (2016). The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood, 127*(20), 2391. doi:10.1182/blood-2016-03-643544

Arthur, J. S. C., & Ley, S. C. (2013). Mitogen-activated protein kinases in innate immunity. *Nature Reviews Immunology, 13*, 679. doi:10.1038/nri3495

Badr, E., ElHefnawi, M., & Heath, L. S. (2016). Computational Identification of Tissue-Specific Splicing Regulatory Elements in Human Genes from RNA-Seq Data. *PLOS ONE, 11*(11), e0166978-e0166978. doi:10.1371/journal.pone.0166978

Baeuerle, P. A. (1995). IκB: a specific inhibitor of the NF-κB transcription factor. *Science, 268*, 522-523.

Baeuerle, P. A., & Baltimore, D. (1991). CHAPTER 20 - The physiology of the NF-κB transcription factor. In *Molecular Aspects of Cellular Regulation* (Vol. 6, pp. 423-446): Elsevier.

Barry, W. T., Nobel, A. B., & Wright, F. A. (2005). Significance analysis of functional categories in gene expression studies: a structured permutation approach. *Bioinformatics (Oxford, England), 21*(9), 1943-1949.

Beißbarth, T., & Speed, T. P. (2004). GOstat: find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics (Oxford, England), 20*(9), 1464-1465.

Berns, K., Hijmans, E. M., Mullenders, J., Brummelkamp, T. R., Velds, A., Heimerikx, M., . . . Weigelt, B. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature, 428*(6981), 431.

Beutler, B. (2004). Inferences, questions and possibilities in Toll-like receptor signalling. *Nature, 430*(6996), 257-263. doi:10.1038/nature02761

Beutler, B., Greenwald, D., Hulmes, J. D., Chang, M., Pan, Y. C., Mathison, J., . . . Cerami, A. (1985). Identity of tumour necrosis factor and the macrophage-secreted factor cachectin. *Nature, 316*(6028), 552-554. doi:10.1038/316552a0

Beutler, B. A., Milsark, I. W., & Cerami, A. (1985). Cachectin/tumor necrosis factor: production, distribution, and metabolic fate in vivo. *The Journal of Immunology, 135*(6), 3972. Retrieved from http://www.jimmunol.org/content/135/6/3972.abstract

Bhinder, B., & Djaballah, H. (2013). Systematic analysis of RNAi reports identifies dismal commonality at gene-level and reveals an unprecedented enrichment in pooled shRNA screens. *Comb Chem High Throughput Screen, 16*(9), 665-681. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/23848309

Bianchi, M. E. (2007). DAMPs, PAMPs and alarmins: all we need to know about danger. *J Leukoc Biol, 81*(1), 1-5. doi:10.1189/jlb.0306164

Birmingham, A., Selfors, L. M., Forster, T., Wrobel, D., Kennedy, C. J., Shanks, E., . . . Shamu, C. E. (2009). Statistical methods for analysis of high-throughput RNA interference screens. *Nature Methods, 6*, 569. doi:10.1038/nmeth.1351

https://www.nature.com/articles/nmeth.1351#supplementary-information

Bjorkbacka, H., Fitzgerald, K. A., Huet, F., Li, X., Gregory, J. A., Lee, M. A., . . . Freeman, M. W. (2004). The induction of macrophage gene expression by LPS predominantly utilizes Myd88-independent signaling cascades. *Physiol Genomics, 19*(3), 319-330. doi:10.1152/physiolgenomics.00128.2004

Black, D. L. (2003). Mechanisms of Alternative Pre-Messenger RNA Splicing. *Annual Review of Biochemistry, 72*(1), 291-336. doi:10.1146/annurev.biochem.72.121801.161720

Blumhagen, R. Z., Hedin, B. R., Malcolm, K. C., Burnham, E. L., Moss, M., Abraham, E., . . . Alper, S. (2017). Alternative pre-mRNA splicing of Toll-like receptor signaling components in peripheral blood mononuclear cells from patients with ARDS. *Am J Physiol Lung Cell Mol Physiol, 313*(5), L930-l939. doi:10.1152/ajplung.00247.2017

Boivin, A., & Mesrobeanu, L. (1935). Recherches sur les antigenes somatiques et sur les endotoxines des bacteries. I. Considerations generales et expose des techniques utilisees. *Rev. immunol, 1*, 553-569.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D(3): Data-Driven Documents. *IEEE Trans Vis Comput Graph, 17*(12), 2301-2309. doi:10.1109/tvcg.2011.185

Boutros, M., & Ahringer, J. (2008). The art and design of genetic screens: RNA interference. *Nature Reviews Genetics, 9*, 554. doi:10.1038/nrg2364

Boutros, M., Kiger, A. A., Armknecht, S., Kerr, K., Hild, M., Koch, B., . . . Heidelberg Fly Array, C. (2004). Genome-wide RNAi analysis of growth and viability in Drosophila cells. *Science, 303*(5659), 832-835.

Brass, A. L., Dykxhoorn, D. M., Benita, Y., Yan, N., Engelman, A., Xavier, R. J., . . . Elledge, S. J. (2008). Identification of host proteins required for HIV infection through a functional genomic screen. *Science, 319*(5865), 921-926. doi:10.1126/science.1152725

Brunet, J. P., Tamayo, P., Golub, T. R., & Mesirov, J. P. (2004). Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A, 101*(12), 4164-4169. doi:10.1073/pnas.0308531101

Bryant, C. E., Symmons, M., & Gay, N. J. (2015). Toll-like receptor signalling through macromolecular protein complexes. *Molecular Immunology, 63*(2), 162-165. doi:https://doi.org/10.1016/j.molimm.2014.06.033

Bürckstümmer, T., Baumann, C., Blüml, S., Dixit, E., Dürnberger, G., Jahn, H., . . . Superti-Furga, G. (2009). An orthogonal proteomic-genomic screen identifies AIM2 as a cytoplasmic DNA sensor for the inflammasome. *Nature Immunology, 10*, 266. doi:10.1038/ni.1702

https://www.nature.com/articles/ni.1702#supplementary-information

Burns, K., Janssens, S., Brissoni, B., Olivos, N., Beyaert, R., & Tschopp, J. (2003). Inhibition of interleukin 1 receptor/Toll-like receptor signaling through the alternatively spliced, short form of MyD88 is due to its failure to recruit IRAK-4. *Journal of Experimental Medicine, 197*(2), 263-268.

Bushman, F. D., Malani, N., Fernandes, J., D'Orso, I., Cagney, G., Diamond, T. L., . . . Chanda, S. K. (2009). Host cell factors in HIV replication: meta-analysis of genome-wide studies. *PLoS Pathog, 5*(5), e1000437. doi:10.1371/journal.ppat.1000437

Carey, M. A., & Papin, J. A. (2018). Ten simple rules for biologists learning to program. *PLOS Computational Biology, 14*(1), e1005871. doi:10.1371/journal.pcbi.1005871

Carmi, Y., Voronov, E., Dotan, S., Lahat, N., Rahat, M. A., Fogel, M., . . . Apte, R. N. (2009). The role of macrophage-derived IL-1 in induction and maintenance of angiogenesis. *J Immunol, 183*(7), 4705-4714. doi:10.4049/jimmunol.0901511

Centanni, E., & Bruschettini, A. (1894). Untersuchungen über das Infectionsfieber. *DMW-Deutsche Medizinische Wochenschrift, 20*(12), 270-272.

Chadban, S. J., Tesch, G. H., Foti, R., Lan, H. Y., Atkins, R. C., & Nikolic-Paterson, D. J. (1998). Interleukin-10 differentially modulates MHC class II expression by mesangial cells and macrophages in vitro and in vivo. *Immunology, 94*(1), 72-78. doi:10.1046/j.1365-2567.1998.00487.x

Chen, L., Paquette, N., Mamoor, S., Rus, F., Nandy, A., Leszyk, J., . . . Silverman, N. (2017). Innate immune signaling in Drosophila is regulated by transforming growth factor beta (TGFbeta)-activated kinase (Tak1)-triggered ubiquitin editing. *J Biol Chem, 292*(21), 8738-8749. doi:10.1074/jbc.M117.788158

Cheng, Z., Taylor, B., Ourthiague, D. R., & Hoffmann, A. (2015). Distinct single-cell signaling characteristics are conferred by the MyD88 and TRIF pathways during TLR4 activation. *Sci Signal, 8*(385), ra69. doi:10.1126/scisignal.aaa5208

Chow, J., Franz, K. M., & Kagan, J. C. (2015). PRRs are watching you: Localization of innate sensing and signaling regulators. *Virology, 479-480*, 104-109. doi:10.1016/j.virol.2015.02.051

Ciechanover, A., Gonen, H., Bercovich, B., Cohen, S., Fajerman, I., Israel, A., . . . Orian, A. (2001). Mechanisms of ubiquitin-mediated, limited processing of the NF-kappaB1 precursor protein p105. *Biochimie, 83*(3-4), 341-349. doi:10.1016/s0300-9084(01)01239-1

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., . . . Zhang, F. (2013). Multiplex genome engineering using CRISPR/Cas systems. *Science, 339*(6121), 819-823. doi:10.1126/science.1231143

Coux, O., & Goldberg, A. L. (1998). Enzymes catalyzing ubiquitination and proteolytic processing of the p105 precursor of nuclear factor kappaB1. *J Biol Chem, 273*(15), 8820-8828. doi:10.1074/jbc.273.15.8820

Cowen, L., Ideker, T., Raphael, B. J., & Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics, 18*(9), 551-562. doi:10.1038/nrg.2017.38

Craven, M. (2015). Gadget. Retrieved from http://gadget.biostat.wisc.edu/

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., … (2015). Pathway and network analysis of cancer genomes. *Nat Methods, 12*(7), 615-621. doi:10.1038/nmeth.3440

Culver-Cochran, A. E., & Starczynowski, D. T. (2018). Chronic innate immune signaling results in ubiquitination of splicing machinery. *Cell Cycle, 17*(4), 407-409. doi:10.1080/15384101.2018.1429082

Dalton, D. K., Pitts-Meek, S., Keshav, S., Figari, I. S., Bradley, A., & Stewart, T. A. (1993). Multiple defects of immune cell function in mice with disrupted interferon-gamma genes. *Science, 259*(5102), 1739. doi:10.1126/science.8456300

Das, M., Sabio, G., Jiang, F., Rincón, M., Flavell, R. A., & Davis, R. J. (2009). Induction of Hepatitis by JNK-Mediated Expression of TNF-α. *Cell, 136*(2), 249-260. doi:https://doi.org/10.1016/j.cell.2008.11.017

De Arras, L., Seng, A., Lackford, B., Keikhaee, M. R., Bowerman, B., Freedman, J. H., . . . Alper, S. (2013). An Evolutionarily Conserved Innate Immunity Protein Interaction Network. *Journal of Biological Chemistry, 288*(3), 1967-1978. Retrieved from http://www.jbc.org/content/288/3/1967.abstract

Decker, T., Müller, M., & Stockinger, S. (2005). The Yin and Yang of type I interferon activity in bacterial infection. *Nature Reviews Immunology, 5*(9), 675-687. doi:10.1038/nri1684

Deng, L., Wang, C., Spencer, E., Yang, L., Braun, A., You, J., . . . Chen, Z. J. (2000). Activation of the IκB Kinase Complex by TRAF6 Requires a Dimeric Ubiquitin-Conjugating Enzyme Complex and a Unique Polyubiquitin Chain. *Cell, 103*(2), 351-361. doi:https://doi.org/10.1016/S0092-8674(00)00126-4

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., . . . Gingeras, T. R. (2013). STAR: ultrafast universal RNA-seq aligner. *Bioinformatics (Oxford, England), 29*(1), 15-21. doi:10.1093/bioinformatics/bts635

Dong, X., Hao, Y., Wang, X., & Tian, W. (2016). LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. *Scientific Reports, 6*(1), 18871. doi:10.1038/srep18871

Dorrington, M. G., & Fraser, I. D. C. (2019). NF-κB Signaling in Macrophages: Dynamics, Crosstalk, and Signal Integration. *Frontiers in immunology, 10*, 705-705. doi:10.3389/fimmu.2019.00705

Doyle, S. L., Husebye, H., Connolly, D. J., Espevik, T., O'Neill, L. A. J., & McGettrick, A. F. (2012). The GOLD domain-containing protein TMED7 inhibits TLR4 signalling from the endosome upon LPS stimulation. *Nature Communications, 3*(1), 707. doi:10.1038/ncomms1706

Durinck, S., Spellman, P. T., Birney, E., & Huber, W. (2009). Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt. *Nat Protoc, 4*(8), 1184-1191. doi:10.1038/nprot.2009.97

Dutta, B., Azhir, A., Merino, L. H., Guo, Y., Revanur, S., Madhamshettiwar, P. B., . . . Fraser, I. D. (2016). An interactive web-based application for Comprehensive Analysis of RNAi-screen Data. *Nat Commun, 7*, 10578. doi:10.1038/ncomms10578

Dutta, B., Wallqvist, A., & Reifman, J. (2012). PathNet: a tool for pathway analysis using topological information. *Source code for biology and medicine, 7*(1), 10-10. doi:10.1186/1751-0473-7-10

Edwards, J. P., Zhang, X., Frauwirth, K. A., & Mosser, D. M. (2006). Biochemical and functional characterization of three activated macrophage populations. *Journal of Leukocyte Biology, 80*(6), 1298-1307. doi:10.1189/jlb.0406249

Effenberger, K. A., Urabe, V. K., & Jurica, M. S. (2017). Modulating splicing with small molecular inhibitors of the spliceosome. *Wiley interdisciplinary reviews. RNA, 8*(2), 10.1002/wrna.1381. doi:10.1002/wrna.1381

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., . . . Turner, S. (2009). Real-Time DNA Sequencing from Single Polymerase Molecules. *Science, 323*(5910), 133. doi:10.1126/science.1162986

Ein-Dor, L., Zuk, O., & Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc Natl Acad Sci U S A, 103*(15), 5923-5928. doi:10.1073/pnas.0601231103

Epelman, S., Lavine, K. J., Beaudin, A. E., Sojka, D. K., Carrero, J. A., Calderon, B., . . . Mann, D. L. (2014). Embryonic and adult-derived resident cardiac macrophages are maintained through distinct mechanisms at steady state and during inflammation. *Immunity, 40*(1), 91-104. doi:10.1016/j.immuni.2013.11.019

Erridge, C., Bennett-Guerrero, E., & Poxton, I. R. (2002). Structure and function of lipopolysaccharides. *Microbes Infect, 4*(8), 837-851. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12270731

Fang, J., Bolanos, L. C., Choi, K., Liu, X., Christie, S., Akunuru, S., . . . Starczynowski, D. T. (2017). Ubiquitination of hnRNPA1 by TRAF6 links chronic innate immune signaling with myelodysplasia. *Nat Immunol, 18*(2), 236-245. doi:10.1038/ni.3654

Fenton, M. J., & Golenbock, D. T. (1998). LPS-binding proteins and receptors. *Journal of Leukocyte Biology, 64*(1), 25-32. doi:10.1002/jlb.64.1.25

Fernandes-Alnemri, T., Yu, J.-W., Datta, P., Wu, J., & Alnemri, E. S. (2009). AIM2 activates the inflammasome and cell death in response to cytoplasmic DNA. *Nature, 458*, 509. doi:10.1038/nature07710

https://www.nature.com/articles/nature07710#supplementary-information

Finkbeiner, S., Frumkin, M., & Kassner, Paul D. (2015). Cell-Based Screening: Extracting Meaning from Complex Data. *Neuron, 86*(1), 160-174. doi:https://doi.org/10.1016/j.neuron.2015.02.023

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature, 391*(6669), 806-811. doi:10.1038/35888

Fisher, L. D. (1993). *Biostatistics: a methodology for the health sciences*. Retrieved from

Fisher, R. A. (1925). *Statistical Methods for Research Workers*: Oliver and Boyd.

Fisher, R. A. (1960). The design of experiments. *The design of experiments.*(7th Ed).

Fitzgerald, K. A., McWhirter, S. M., Faia, K. L., Rowe, D. C., Latz, E., Golenbock, D. T., . . . Maniatis, T. (2003). IKKε and TBK1 are essential components of the IRF3 signaling pathway. *Nature Immunology, 4*(5), 491-496. doi:10.1038/ni921

Fitzgerald, K. A., Palsson-McDermott, E. M., Bowie, A. G., Jefferies, C. A., Mansell, A. S., Brady, G., . . . Harte, M. T. (2001). Mal (MyD88-adapter-like) is required for Toll-like receptor-4 signal transduction. *Nature, 413*(6851), 78.

Fitzgerald, K. A., Rowe, D. C., Barnes, B. J., Caffrey, D. R., Visintin, A., Latz, E., . . . Golenbock, D. T. (2003). LPS-TLR4 signaling to IRF-3/7 and NF-kappaB involves the toll adapters TRAM and TRIF. *J Exp Med, 198*(7), 1043-1055. doi:10.1084/jem.20031023

Franchi, L., Muñoz-Planillo, R., & Núñez, G. (2012). Sensing and reacting to microbes through the inflammasomes. *Nature Immunology, 13*(4), 325-332. doi:10.1038/ni.2231

Fritz, J. H., Ferrero, R. L., Philpott, D. J., & Girardin, S. E. (2006). Nod-like proteins in immunity, inflammation and disease. *Nature Immunology, 7*(12), 1250-1257. doi:10.1038/ni1412

Fury, W., Batliwalla, F., Gregersen, P. K., & Li, W. (2006). Overlapping probabilities of top ranking gene lists, hypergeometric distribution, and stringency of gene selection criterion. *Conf Proc IEEE Eng Med Biol Soc, 1*, 5531-5534. doi:10.1109/IEMBS.2006.260828

Gay, N. J., Gangloff, M., & O'Neill, L. A. J. (2011). What the Myddosome structure tells us about the initiation of innate immunity. *Trends in Immunology, 32*(3), 104-109. doi:https://doi.org/10.1016/j.it.2010.12.005

Gay, N. J., & Keith, F. J. (1991). Drosophila Toll and IL-1 receptor. *Nature, 351*(6325), 355-356. doi:10.1038/351355b0

Ge, Y., & Porse, B. T. (2014). The functional consequences of intron retention: Alternative splicing coupled to NMD as a regulator of gene expression. *BioEssays, 36*(3), 236-243. doi:10.1002/bies.201300156

Geissmann, F., Jung, S., & Littman, D. R. (2003). Blood Monocytes Consist of Two Principal Subsets with Distinct Migratory Properties. *Immunity, 19*(1), 71-82. doi:https://doi.org/10.1016/S1074-7613(03)00174-2

Geistlinger, L., Csaba, G., Santarelli, M., Ramos, M., Schiffer, L., Turaga, N., . . . Waldron, L. (2020). Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in Bioinformatics*. doi:10.1093/bib/bbz158

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., . . . Zhang, J. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol, 5*(10), R80. doi:10.1186/gb-2004-5-10-r80

Ghosh, S., Gifford, A. M., Riviere, L. R., Tempst, P., Nolan, G. P., & Baltimore, D. (1990). Cloning of the p50 DNA binding subunit of NF-κB: Homology to rel and dorsal. *Cell, 62*(5), 1019-1029. doi:https://doi.org/10.1016/0092-8674(90)90276-K

Ghosh, S., May, M. J., & Kopp, E. B. (1998). NF-κB AND REL PROTEINS: Evolutionarily Conserved Mediators of Immune Responses. *Annual Review of Immunology, 16*(1), 225-260. doi:10.1146/annurev.immunol.16.1.225

Godec, J., Tan, Y., Liberzon, A., Tamayo, P., Bhattacharya, S., Butte, Atul J., . . . Haining, W. N. (2016). Compendium of Immune Signatures Identifies Conserved and Species-Specific Biology in Response to Inflammation. *Immunity, 44*(1), 194-206. doi:https://doi.org/10.1016/j.immuni.2015.12.006

Goeman, J. J., & Bühlmann, P. (2007). Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics, 23*(8), 980-987.

Gordon, S. (2016). Phagocytosis: The Legacy of Metchnikoff. *Cell, 166*(5), 1065-1068. doi:https://doi.org/10.1016/j.cell.2016.08.017

Gottipati, S., Rao, N. L., & Fung-Leung, W.-P. (2008). IRAK1: A critical signaling mediator of innate immunity. *Cellular Signalling, 20*(2), 269-276. doi:https://doi.org/10.1016/j.cellsig.2007.08.009

Gottschalk, R. A., Dorrington, M. G., Dutta, B., Krauss, K. S., Martins, A. J., Uderhardt, S., . . . Germain, R. N. (2019). IFN-mediated negative feedback supports bacteria class-specific macrophage inflammatory responses. *eLife, 8*, e46836. doi:10.7554/eLife.46836

Gray, P., Michelsen, K. S., Sirois, C. M., Lowe, E., Shimada, K., Crother, T. R., . . . Arditi, M. (2010). Identification of a Novel Human MD-2 Splice Variant That Negatively Regulates Lipopolysaccharide-Induced TLR4 Signaling. *The Journal of Immunology, 184*(11), 6359. doi:10.4049/jimmunol.0903543

Grohar, P. J., Kim, S., Rangel Rivera, G. O., Sen, N., Haddock, S., Harlow, M. L., . . . Caplen, N. J. (2016). Functional Genomic Screening Reveals Splicing of the EWS-FLI1 Fusion Transcript as a Vulnerability in Ewing Sarcoma. *Cell Rep, 14*(3), 598-610. doi:10.1016/j.celrep.2015.12.063

Grosso, A. R., Gomes, A. Q., Barbosa-Morais, N. L., Caldeira, S., Thorne, N. P., Grech, G., . . . Carmo-Fonseca, M. (2008). Tissue-specific splicing factor gene expression signatures. *Nucleic Acids Research, 36*(15), 4823-4832. doi:10.1093/nar/gkn463

Gu, Z., Liu, J., Cao, K., Zhang, J., & Wang, J. (2012). Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. *BMC systems biology, 6*(1), 56.

Guha, M., & Mackman, N. (2001). LPS induction of gene expression in human monocytes. *Cellular Signalling, 13*(2), 85-94. doi:https://doi.org/10.1016/S0898-6568(00)00149-2

Guilliams, M., De Kleer, I., Henri, S., Post, S., Vanhoutte, L., De Prijck, S., . . . Lambrecht, B. N. (2013). Alveolar macrophages develop from fetal monocytes that differentiate into long-lived cells in the first week of life via GM-CSF. *J Exp Med, 210*(10), 1977-1992. doi:10.1084/jem.20131199

Guilliams, M., Mildner, A., & Yona, S. (2018). Developmental and Functional Heterogeneity of Monocytes. *Immunity, 49*(4), 595-613. doi:10.1016/j.immuni.2018.10.005

Hagar, J. A., Powell, D. A., Aachoui, Y., Ernst, R. K., & Miao, E. A. (2013). Cytoplasmic LPS Activates Caspase-11: Implications in TLR4-Independent Endotoxic Shock. *Science, 341*(6151), 1250. doi:10.1126/science.1240988

Hammad, H., & Lambrecht, B. N. (2011). Dendritic cells and airway epithelial cells at the interface between innate and adaptive immune responses. *Allergy, 66*(5), 579-587. doi:10.1111/j.1398-9995.2010.02528.x

Hao, L., He, Q., Wang, Z., Craven, M., Newton, M. A., & Ahlquist, P. (2013). Limited Agreement of Independent RNAi Screens for Virus-Required Host Genes Owes More to False-Negative than False-Positive Factors. *PLOS Computational Biology, 9*(9), e1003235. doi:10.1371/journal.pcbi.1003235

Hartmannová, H., Piherová, L., Tauchmannová, K., Kidd, K., Acott, P. D., Crocker, J. F. S., . . . Kmoch, S. (2016). Acadian variant of Fanconi syndrome is caused by mitochondrial respiratory chain complex I deficiency due to a non-coding mutation in complex I assembly factor NDUFAF6. *Human Molecular Genetics, 25*(18), 4062-4079. doi:10.1093/hmg/ddw245

Hashimoto, C., Hudson, K. L., & Anderson, K. V. (1988). The Toll gene of Drosophila, required for dorsal-ventral embryonic polarity, appears to encode a transmembrane protein. *Cell, 52*(2), 269-279.

Heil, F., Hemmi, H., Hochrein, H., Ampenberger, F., Kirschning, C., Akira, S., . . . Bauer, S. (2004). Species-specific recognition of single-stranded RNA via toll-like receptor 7 and 8. *Science, 303*(5663), 1526-1529. doi:10.1126/science.1093620

Hirsch, A. J. (2010). The use of RNAi-based screens to identify host proteins involved in viral replication. *Future microbiology, 5*(2), 303-311.

Hoffmann, J. A. (2003). The immune response of Drosophila. *Nature, 426*(6962), 33-38. doi:10.1038/nature02021

Holten, D. (2006). Hierarchical Edge Bundles: Visualization of Adjacency Relations in Hierarchical Data. *IEEE Trans Vis Comput Graph, 12*(5), 741-748. doi:10.1109/TVCG.2006.147

Hornung, V., Ablasser, A., Charrel-Dennis, M., Bauernfeind, F., Horvath, G., Caffrey, D. R., . . . Fitzgerald, K. A. (2009). AIM2 recognizes cytosolic dsDNA and forms a caspase-1-activating inflammasome with ASC. *Nature, 458*, 514. doi:10.1038/nature07725
https://www.nature.com/articles/nature07725#supplementary-information

Hornung, V., Ellegast, J., Kim, S., Brzózka, K., Jung, A., Kato, H., . . . Hartmann, G. (2006). 5'-Triphosphate RNA Is the Ligand for RIG-I. *Science, 314*(5801), 994. doi:10.1126/science.1132505

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2008). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols, 4*, 44. doi:10.1038/nprot.2008.211
https://www.nature.com/articles/nprot.2008.211#supplementary-information
Hurst, S. M., Wilkinson, T. S., McLoughlin, R. M., Jones, S., Horiuchi, S., Yamamoto, N., . . . Jones, S. A. (2001). IL-6 and Its Soluble Receptor Orchestrate a Temporal Switch in the Pattern of Leukocyte Recruitment Seen during Acute Inflammation. *Immunity, 14*(6), 705-714. doi:https://doi.org/10.1016/S1074-7613(01)00151-0
Infantino, V., Convertini, P., Cucci, L., Panaro, M. A., Di Noia, M. A., Calvello, R., . . . Iacobazzi, V. (2011). The mitochondrial citrate carrier: a new player in inflammation. *Biochemical Journal, 438*(3), 433-436.
Ishikawa, H., & Barber, G. N. (2008). STING is an endoplasmic reticulum adaptor that facilitates innate immune signalling. *Nature, 455*(7213), 674-678. doi:10.1038/nature07317
Ishikawa, H., Ma, Z., & Barber, G. N. (2009). STING regulates intracellular DNA-mediated, type I interferon-dependent innate immunity. *Nature, 461*(7265), 788-792. doi:10.1038/nature08476
Iwami, K.-i., Matsuguchi, T., Masuda, A., Kikuchi, T., Musikacharoen, T., & Yoshikai, Y. (2000). Cutting edge: naturally occurring soluble form of mouse Toll-like receptor 4 inhibits lipopolysaccharide signaling. *The Journal of Immunology, 165*(12), 6682-6686.
Jackson, A. L., & Linsley, P. S. (2004). Noise amidst the silence: off-target effects of siRNAs? *Trends Genet, 20*(11), 521-524. doi:10.1016/j.tig.2004.08.006
Jackson, A. L., & Linsley, P. S. (2010). Recognizing and avoiding siRNA off-target effects for target identification and therapeutic application. *Nature Reviews Drug Discovery, 9*(1), 57-67. doi:10.1038/nrd3010
Jacob, A. G., & Smith, C. W. J. (2017). Intron retention as a component of regulated gene expression programs. *Human genetics, 136*(9), 1043-1057. doi:10.1007/s00439-017-1791-x
Janeway, C. A. (1992). The immune system evolved to discriminate infectious nonself from noninfectious self. *Immunology Today, 13*(1), 11-16. doi:https://doi.org/10.1016/0167-5699(92)90198-G
Janeway, C. A., Jr. (1989). Approaching the asymptote? Evolution and revolution in immunology. *Cold Spring Harb Symp Quant Biol, 54 Pt 1*, 1-13. doi:10.1101/sqb.1989.054.01.003
Janeway, C. A., Travers, P., Walport, M., & Shlomchik, M. (1996). *Immunobiology: the immune system in health and disease* (Vol. 7): Current Biology London.
Janssens, S., Burns, K., Tschopp, J., & Beyaert, R. (2002). Regulation of Interleukin-1- and Lipopolysaccharide-Induced NF-κB Activation by Alternative Splicing of MyD88. *Current Biology, 12*(6), 467-471. doi:https://doi.org/10.1016/S0960-9822(02)00712-1
Jarešová, I., Rožková, D., Špíšek, R., Janda, A., Brázová, J., & Šedivá, A. (2007). Kinetics of Toll-like receptor-4 splice variants expression in lipopolysaccharide-stimulated antigen presenting cells of healthy donors and patients with cystic fibrosis. *Microbes and infection, 9*(11), 1359-1367.
Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science, 337*(6096), 816-821. doi:10.1126/science.1225829
John, S. P., Sun, J., Carlson, R. J., Cao, B., Bradfield, C. J., Song, J., . . . Fraser, I. D. C. (2018). IFIT1 Exerts Opposing Regulatory Effects on the Inflammatory and

Interferon Gene Programs in LPS-Activated Human Macrophages. *Cell Rep, 25*(1), 95-106.e106. doi:10.1016/j.celrep.2018.09.002
Joshi-Tope, G., Vastrik, I., Gopinath, G., Matthews, L., Schmidt, E., Gillespie, M., . . . Wu, G. (2003). *The Genome Knowledgebase: a resource for biologists and bioinformaticists.* Paper presented at the Cold Spring Harb Symp Quant Biol.
Kagan, J. C., Su, T., Horng, T., Chow, A., Akira, S., & Medzhitov, R. (2008). TRAM couples endocytosis of Toll-like receptor 4 to the induction of interferon-β. *Nature Immunology, 9*, 361. doi:10.1038/ni1569
https://www.nature.com/articles/ni1569#supplementary-information
Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., & Morishima, K. (2017). KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res, 45*(D1), D353-d361. doi:10.1093/nar/gkw1092
Kanehisa, M., & Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res, 28*(1), 27-30. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10592173
Kaplow, I. M., Singh, R., Friedman, A., Bakal, C., Perrimon, N., & Berger, B. (2009). RNAiCut: automated detection of significant genes from functional genomic screens. *Nature Methods, 6*(7), 476-477. doi:10.1038/nmeth0709-476
Kato, H., Takeuchi, O., Sato, S., Yoneyama, M., Yamamoto, M., Matsui, K., . . . Akira, S. (2006). Differential roles of MDA5 and RIG-I helicases in the recognition of RNA viruses. *Nature, 441*(7089), 101-105. doi:10.1038/nature04734
Kayagaki, N., Wong, M. T., Stowe, I. B., Ramani, S. R., Gonzalez, L. C., Akashi-Takamura, S., . . . Dixit, V. M. (2013). Noncanonical Inflammasome Activation by Intracellular LPS Independent of TLR4. *Science, 341*(6151), 1246. doi:10.1126/science.1240248
Keating, S. E., Maloney, G. M., Moran, E. M., & Bowie, A. G. (2007). IRAK-2 Participates in Multiple Toll-like Receptor Signaling Pathways to NFκB via Activation of TRAF6 Ubiquitination. *Journal of Biological Chemistry, 282*(46), 33435-33443. Retrieved from http://www.jbc.org/content/282/46/33435
KEGG, K. E. o. G. a. G. Toll-like receptor signaling pathway - Homo sapiens (human). Retrieved from https://www.genome.jp/dbget-bin/www_bget?pathway+hsa04620
Kelly, B., & O'Neill, L. A. J. (2015). Metabolic reprogramming in macrophages and dendritic cells in innate immunity. *Cell Research, 25*, 771. doi:10.1038/cr.2015.68
Khatri, P., Sirota, M., & Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLOS Computational Biology, 8*(2), e1002375. doi:10.1371/journal.pcbi.1002375
Kim, C., Sano, Y., Todorova, K., Carlson, B. A., Arpa, L., Celada, A., . . . Park, J. M. (2008). The kinase p38α serves cell type–specific inflammatory functions in skin injury and coordinates pro- and anti-inflammatory gene expression. *Nature Immunology, 9*(9), 1019-1027. doi:10.1038/ni.1640
Kimbrell, D. A., & Beutler, B. (2001). The evolution and genetics of innate immunity. *Nat Rev Genet, 2*(4), 256-267. doi:10.1038/35066006
Kniepert, A., & Groettrup, M. (2014). The unique functions of tissue-specific proteasomes. *Trends Biochem Sci, 39*(1), 17-24. doi:10.1016/j.tibs.2013.10.004
König, R., Zhou, Y., Elleder, D., Diamond, T. L., Bonamy, G. M. C., Irelan, J. T., . . . Chanda, S. K. (2008). Global Analysis of Host-Pathogen Interactions that Regulate Early-Stage HIV-1 Replication. *Cell, 135*(1), 49-60. doi:https://doi.org/10.1016/j.cell.2008.07.032
Kovalenko, A., Chable-Bessia, C., Cantarella, G., Israël, A., Wallach, D., & Courtois, G. (2003). The tumour suppressor CYLD negatively regulates NF-κB signalling by deubiquitination. *Nature, 424*(6950), 801.

Krawczyk, C. M., Holowka, T., Sun, J., Blagih, J., Amiel, E., DeBerardinis, R. J., . . . Pearce, E. J. (2010). Toll-like receptor-induced changes in glycolytic metabolism regulate dendritic cell activation. *Blood, 115*(23), 4742-4749. doi:10.1182/blood-2009-10-249540

Laird, M. H. W., Rhee, S. H., Perkins, D. J., Medvedev, A. E., Piao, W., Fenton, M. J., & Vogel, S. N. (2009). TLR4/MyD88/PI3K interactions regulate TLR4 signaling. *Journal of Leukocyte Biology, 85*(6), 966-977. doi:10.1189/jlb.1208763

Langfelder, P., & Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol, 1*, 54. doi:10.1186/1752-0509-1-54

Lawrence, T., & Natoli, G. (2011). Transcriptional regulation of macrophage polarization: enabling diversity with identity. *Nature Reviews Immunology, 11*, 750. doi:10.1038/nri3088

Lee, Y., & Rio, D. C. (2015). Mechanisms and Regulation of Alternative Pre-mRNA Splicing. *Annual Review of Biochemistry, 84*, 291-323. doi:10.1146/annurev-biochem-060614-034316

Lemaitre, B., Nicolas, E., Michaut, L., Reichhart, J.-M., & Hoffmann, J. A. (1996). The Dorsoventral Regulatory Gene Cassette spätzle/Toll/cactus Controls the Potent Antifungal Response in Drosophila Adults. *Cell, 86*(6), 973-983. doi:https://doi.org/10.1016/S0092-8674(00)80172-5

Li, N., Katz, S., Dutta, B., Benet, Z. L., Sun, J., & Fraser, I. D. C. (2017). Genome-wide siRNA screen of genes regulating the LPS-induced NF-κB and TNF-α responses in mouse macrophages. *4*, 170008. doi:10.1038/sdata.2017.8

https://www.nature.com/articles/sdata20178#supplementary-information

Li, N., Sun, J., Benet, Z. L., Wang, Z., Al-Khodor, S., John, S. P., . . . Fraser, I. D. (2015). Development of a cell system for siRNA screening of pathogen responses in human and mouse macrophages. *Sci Rep, 5*, 9559. doi:10.1038/srep09559

Li, X., Shu, C., Yi, G., Chaton, C. T., Shelton, C. L., Diao, J., . . . Li, P. (2013). Cyclic GMP-AMP synthase is activated by double-stranded DNA-induced oligomerization. *Immunity, 39*(6), 1019-1031.

Liew, F. Y., Xu, D., Brint, E. K., & O'Neill, L. A. J. (2005). Negative regulation of Toll-like receptor-mediated immune responses. *Nature Reviews Immunology, 5*(6), 446-458. doi:10.1038/nri1630

Lin, J., Hu, Y., Nunez, S., Foulkes Andrea, S., Cieply, B., Xue, C., . . . Reilly Muredach, P. (2016). Transcriptome-Wide Analysis Reveals Modulation of Human Macrophage Inflammatory Phenotype Through Alternative Splicing. *Arteriosclerosis, Thrombosis, and Vascular Biology, 36*(7), 1434-1447. doi:10.1161/ATVBAHA.116.307573

Liu, A., Gong, P., Hyun, S. W., Wang, K. Z. Q., Cates, E. A., Perkins, D., . . . Goldblum, S. E. (2012). TRAF6 Protein Couples Toll-like Receptor 4 Signaling to Src Family Kinase Activation and Opening of Paracellular Pathway in Human Lung Microvascular Endothelia. *Journal of Biological Chemistry, 287*(20), 16132-16145. Retrieved from http://www.jbc.org/content/287/20/16132

Liu, H., Lorenzini, P. A., Zhang, F., Xu, S., Wong, Mei S. M., Zheng, J., & Roca, X. (2018). Alternative splicing analysis in human monocytes and macrophages reveals MBNL1 as major regulator. *Nucleic Acids Research, 46*(12), 6069-6086. doi:10.1093/nar/gky401

Lotterhos, K. E., François, O., & Blum, M. G. B. (2016). Not just methods: User expertise explains the variability of outcomes of genome-wide studies. *bioRxiv*, 055046. doi:10.1101/055046

Lüderitz, O., Westphal, O., Staub, A. M., & Nikaido, H. (1971). Microbial Toxins (WEINBAUM, G., KADIS, S. and AJL, SJ eds). Vol. IV. In: Academic Press, New York.

Ma, J., Wang, J., Ghoraie, L. S., Men, X., Haibe-Kains, B., & Dai, P. (2019). A Comparative Study of Cluster Detection Algorithms in Protein-Protein Interaction for Drug Target Discovery and Drug Repurposing. *Frontiers in pharmacology, 10*, 109-109. doi:10.3389/fphar.2019.00109

Mahotka, C., Liebmann, J., Wenzel, M., Suschek, C. V., Schmitt, M., Gabbert, H. E., & Gerharz, C. D. (2002). Differential subcellular localization of functionally divergent survivin splice variants. *Cell Death & Differentiation, 9*(12), 1334-1342. doi:10.1038/sj.cdd.4401091

Malo, N., Hanley, J. A., Cerquozzi, S., Pelletier, J., & Nadon, R. (2006). Statistical practice in high-throughput screening data analysis. *Nature Biotechnology, 24*(2), 167-175. doi:10.1038/nbt1186

Manly, K. F., Nettleton, D., & Hwang, J. T. G. (2004). Genomics, Prior Probability, and Statistical Tests of Multiple Hypotheses. *Genome Research, 14*(6), 997-1001. Retrieved from http://genome.cshlp.org/content/14/6/997.short

Marine, S., Bahl, A., Ferrer, M., & Buehler, E. (2012). Common seed analysis to identify off-target effects in siRNA screens. *J Biomol Screen, 17*(3), 370-378.

Martin, M., Schifferle, R. E., Cuesta, N., Vogel, S. N., Katz, J., & Michalek, S. M. (2003). Role of the Phosphatidylinositol 3 Kinase-Akt Pathway in the Regulation of IL-10 and IL-12 by Porphyromonas gingivalis Lipopolysaccharide. *The Journal of Immunology, 171*(2), 717. doi:10.4049/jimmunol.171.2.717

Mathur, R., Rotroff, D., Ma, J., Shojaie, A., & Motsinger-Reif, A. (2018). Gene set analysis methods: a systematic comparison. *BioData Mining, 11*(1), 8. doi:10.1186/s13040-018-0166-8

Matzinger, P. (1994). Tolerance, danger, and the extended family. *Annu Rev Immunol, 12*, 991-1045. doi:10.1146/annurev.iy.12.040194.005015

Mayeda, A., Screaton, G. R., Chandler, S. D., Fu, X. D., & Krainer, A. R. (1999). Substrate specificities of SR proteins in constitutive splicing are determined by their RNA recognition motifs and composite pre-mRNA exonic elements. *Mol Cell Biol, 19*(3), 1853-1863. doi:10.1128/mcb.19.3.1853

Medzhitov, R., Preston-Hurlburt, P., & Janeway, C. A. (1997). A human homologue of the Drosophila Toll protein signals activation of adaptive immunity. *Nature, 388*(6640), 394-397. doi:10.1038/41131

Medzhitov, R., Preston-Hurlburt, P., Kopp, E., Stadlen, A., Chen, C., Ghosh, S., & Janeway Jr, C. A. (1998). MyD88 is an adaptor protein in the hToll/IL-1 receptor family signaling pathways. *Molecular cell, 2*(2), 253-258.

Mertins, P., Przybylski, D., Yosef, N., Qiao, J., Clauser, K., Raychowdhury, R., . . . Chevrier, N. (2017). An Integrative Framework Reveals Signaling-to-Transcription Events in Toll-like Receptor Signaling. *Cell Rep, 19*(13), 2853-2866. doi:10.1016/j.celrep.2017.06.016

Michelucci, A., Cordes, T., Ghelfi, J., Pailot, A., Reiling, N., Goldmann, O., . . . Rausell, A. (2013). Immune-responsive gene 1 protein links metabolism to immunity by catalyzing itaconic acid production. *Proceedings of the National Academy of Sciences, 110*(19), 7820-7825.

Mills, C. D., Kincaid, K., Alt, J. M., Heilman, M. J., & Hill, A. M. (2000). M-1/M-2 macrophages and the Th1/Th2 paradigm. *J Immunol, 164*(12), 6166-6173. doi:10.4049/jimmunol.164.12.6166

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., . . . Draghici, S. (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology, 4*(278). doi:10.3389/fphys.2013.00278

Modrek, B., & Lee, C. (2002). A genomic view of alternative splicing. *Nature Genetics, 30*(1), 13-19. doi:10.1038/ng0102-13

Mohr, S., Bakal, C., & Perrimon, N. (2010). Genomic Screening with RNAi: Results and Challenges. *Annual Review of Biochemistry, 79*(1), 37-64. doi:10.1146/annurev-biochem-060408-092949

Mooney, M. A., & Wilmot, B. (2015). Gene set analysis: A step-by-step guide. *Am J Med Genet B Neuropsychiatr Genet, 168*(7), 517-527. doi:10.1002/ajmg.b.32328

Morozov, A. V., & Karpov, V. L. (2019). Proteasomes and Several Aspects of Their Heterogeneity Relevant to Cancer. *Frontiers in Oncology, 9*, 761. Retrieved from https://www.frontiersin.org/article/10.3389/fonc.2019.00761

Mosser, D. M., & Edwards, J. P. (2008). Exploring the full spectrum of macrophage activation. *Nature Reviews Immunology, 8*, 958. doi:10.1038/nri2448

Murray, P. J., & Wynn, T. A. (2011). Protective and pathogenic functions of macrophage subsets. *Nat Rev Immunol, 11*(11), 723-737. doi:10.1038/nri3073

Muzio, M., Natoli, G., Saccani, S., Levrero, M., & Mantovani, A. (1998). The human Toll signaling pathway: divergence of nuclear factor κB and JNK/SAPK activation upstream of tumor necrosis factor receptor–associated factor 6 (TRAF6). *Journal of Experimental Medicine, 187*(12), 2097-2101.

Nakao, H., Ukai, I., & Kotani, J. (2017). A review of the history of the origin of triage from a disaster medicine perspective. *Acute medicine & surgery, 4*(4), 379-384. doi:10.1002/ams2.293

Nau, G. J., Richmond, J. F. L., Schlesinger, A., Jennings, E. G., Lander, E. S., & Young, R. A. (2002). Human macrophage activation programs induced by bacterial pathogens. *Proceedings of the National Academy of Sciences, 99*(3), 1503. doi:10.1073/pnas.022649799

Netea, M. G., Gow, N. A. R., Munro, C. A., Bates, S., Collins, C., Ferwerda, G., . . . Kullberg, B. J. (2006). Immune sensing of Candida albicans requires cooperative recognition of mannans and glucans by lectin and Toll-like receptors. *The Journal of Clinical Investigation, 116*(6), 1642-1650. doi:10.1172/JCI27114

Netea, M. G., Joosten, L. A. B., Latz, E., Mills, K. H. G., Natoli, G., Stunnenberg, H. G., . . . Xavier, R. J. (2016). Trained immunity: A program of innate immune memory in health and disease. *Science, 352*(6284), aaf1098. doi:10.1126/science.aaf1098

Nguyen, T., Tagett, R., Donato, M., Mitrea, C., & Draghici, S. (2015). A novel bi-level meta-analysis approach: applied to biological pathway analysis. *Bioinformatics, 32*(3), 409-416. doi:10.1093/bioinformatics/btv588

Nguyen, T.-M., Shafi, A., Nguyen, T., & Draghici, S. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. *Genome Biol, 20*(1), 203. doi:10.1186/s13059-019-1790-4

Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature, 463*, 457. doi:10.1038/nature08909

Núñez Miguel, R., Wong, J., Westoll, J. F., Brooks, H. J., O'Neill, L. A. J., Gay, N. J., . . . Monie, T. P. (2007). A Dimer of the Toll-Like Receptor 4 Cytoplasmic Domain Provides a Specific Scaffold for the Recruitment of Signalling Adaptor Proteins. *PLOS ONE, 2*(8), e788. doi:10.1371/journal.pone.0000788

O'Donnell, M. A., Perez-Jimenez, E., Oberst, A., Ng, A., Massoumi, R., Xavier, R., . . . Ting, A. T. (2011). Caspase 8 inhibits programmed necrosis by processing CYLD. *Nature cell biology, 13*(12), 1437.

Oka, Y., Varmark, H., Vitting-Seerup, K., Beli, P., Waage, J., Hakobyan, A., . . . Mailand, N. (2014). UBL5 is essential for pre-mRNA splicing and sister chromatid cohesion in human cells. *EMBO reports, 15*(9), 956-964. doi:10.15252/embr.201438679

Ortiz-Lazareno, P. C., Hernandez-Flores, G., Dominguez-Rodriguez, J. R., Lerma-Diaz, J. M., Jave-Suarez, L. F., Aguilar-Lemarroy, A., . . . Bravo-Cuellar, A. (2008). MG132 proteasome inhibitor modulates proinflammatory cytokines production and expression of their receptors in U937 cells: involvement of nuclear factor-kappaB and activator protein-1. *Immunology, 124*(4), 534-541. doi:10.1111/j.1365-2567.2008.02806.x

Oshiumi, H., Matsumoto, M., Funami, K., Akazawa, T., & Seya, T. (2003). TICAM-1, an adaptor molecule that participates in Toll-like receptor 3–mediated interferon-β induction. *Nature Immunology, 4*(2), 161-167. doi:10.1038/ni886

Ostuni, R., Zanoni, I., & Granucci, F. (2010). Deciphering the complexity of Toll-like receptor signaling. *Cellular and Molecular Life Sciences, 67*(24), 4109-4134. doi:10.1007/s00018-010-0464-x

Oti, M., Snel, B., Huynen, M. A., & Brunner, H. G. (2006). Predicting disease genes using protein-protein interactions. *Journal of medical genetics, 43*(8), 691-698. doi:10.1136/jmg.2006.041376

Pan, Q., Bakowski, M. A., Morris, Q., Zhang, W., Frey, B. J., Hughes, T. R., & Blencowe, B. J. (2005). Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends in Genetics, 21*(2), 73-77. doi:https://doi.org/10.1016/j.tig.2004.12.004

Papaemmanuil, E., Gerstung, M., Malcovati, L., Tauro, S., Gundem, G., Van Loo, P., . . . Campbell, P. J. (2013). Clinical and biological implications of driver mutations in myelodysplastic syndromes. *Blood, 122*(22), 3616. doi:10.1182/blood-2013-08-518886

Papasaikas, P., Tejedor, J. R., Vigevani, L., & Valcarcel, J. (2015). Functional splicing network reveals extensive regulatory potential of the core spliceosomal machinery. *Mol Cell, 57*(1), 7-22. doi:10.1016/j.molcel.2014.10.030

Park, B. S., Song, D. H., Kim, H. M., Choi, B.-S., Lee, H., & Lee, J.-O. (2009). The structural basis of lipopolysaccharide recognition by the TLR4–MD-2 complex. *Nature, 458*, 1191. doi:10.1038/nature07830
https://www.nature.com/articles/nature07830#supplementary-information

Parnas, O., Jovanovic, M., Eisenhaure, T. M., Herbst, R. H., Dixit, A., Ye, C. J., . . . Regev, A. (2015). A Genome-wide CRISPR Screen in Primary Immune Cells to Dissect Regulatory Networks. *Cell, 162*(3), 675-686. doi:10.1016/j.cell.2015.06.059

Pawellek, A., McElroy, S., Samatov, T., Mitchell, L., Woodland, A., Ryder, U., . . . Lamond, A. I. (2014). Identification of small molecule inhibitors of pre-mRNA splicing. *J Biol Chem, 289*(50), 34683-34698. doi:10.1074/jbc.M114.590976

Paz, I., Kosti, I., Ares, M., Jr., Cline, M., & Mandel-Gutfreund, Y. (2014). RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Research, 42*(Web Server issue), W361-W367. doi:10.1093/nar/gku406

Perry, G., Iragavarapu-Charyulu, V., Harhaj, E. W., & Torroella-Kouri, M. (2010). Role of the proteasome in the downregulation of transcription factors NFkappaB and C/EBP in macrophages from tumor hosts. *Oncol Rep, 23*(3), 875-881.

Pfeiffer, R. (1892). Untersuchungen über das Choleragift. *Zeitschrift für Hygiene und Infektionskrankheiten, 11*(1), 393-412. doi:10.1007/BF02284303

Pita-Juárez, Y., Altschuler, G., Kariotis, S., Wei, W., Koler, K., Green, C., . . . Hide, W. (2018). The Pathway Coexpression Network: Revealing pathway relationships. *PLOS Computational Biology, 14*(3), e1006042. doi:10.1371/journal.pcbi.1006042

Poltorak, A., He, X., Smirnova, I., Liu, M.-Y., Van Huffel, C., Du, X., . . . Galanos, C. (1998). Defective LPS signaling in C3H/HeJ and C57BL/10ScCr mice: mutations in Tlr4 gene. *Science, 282*(5396), 2085-2088.

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Raetz, C. R. H., & Whitfield, C. (2002). Lipopolysaccharide Endotoxins. *Annual Review of Biochemistry, 71*(1), 635-700. doi:10.1146/annurev.biochem.71.110601.135414

Rahnenführer, J., Domingues, F. S., Maydt, J., & Lengauer, T. (2004). Calculating the statistical significance of changes in pathway activity from gene expression data. *Statistical applications in genetics and molecular biology, 3*(1), 1-29.

Ramanan, V. K., Shen, L., Moore, J. H., & Saykin, A. J. (2012). Pathway analysis of genomic data: concepts, methods, and prospects for future development. *Trends in Genetics, 28*(7), 323-332. doi:https://doi.org/10.1016/j.tig.2012.03.004

Ramasamy, A., Mondry, A., Holmes, C. C., & Altman, D. G. (2008). Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med, 5*(9), e184. doi:10.1371/journal.pmed.0050184

Rao, N., Nguyen, S., Ngo, K., & Fung-Leung, W. P. (2005). A novel splice variant of interleukin-1 receptor (IL-1R)-associated kinase 1 plays a negative regulatory role in Toll/IL-1R-induced inflammatory signaling. *Mol Cell Biol, 25*(15), 6521-6532. doi:10.1128/mcb.25.15.6521-6532.2005

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., & Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Res, 62*(15), 4427-4433. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/12154050

Roder, J., Linstid, B., & Oliveira, C. (2019). Improving the power of gene set enrichment analyses. *BMC Bioinformatics, 20*(1), 257. doi:10.1186/s12859-019-2850-1

Rodríguez-García, E., Olagüe, C., Ríus-Rocabert, S., Ferrero, R., Llorens, C., Larrea, E., . . . Nistal-Villan, E. (2018). TMEM173 Alternative Spliced Isoforms Modulate Viral Replication through the STING Pathway. *ImmunoHorizons, 2*(11), 363. doi:10.4049/immunohorizons.1800068

Rosenbluh, J., Mercer, J., Shrestha, Y., Oliver, R., Tamayo, P., Doench, John G., . . . Hahn, William C. (2016). Genetic and Proteomic Interrogation of Lower Confidence Candidate Genes Reveals Signaling Networks in β-Catenin-Active Cancers. *Cell Systems, 3*(3), 302-316.e304. doi:https://doi.org/10.1016/j.cels.2016.09.001

Sakai, J., Cammarota, E., Wright, J. A., Cicuta, P., Gottschalk, R. A., Li, N., . . . Bryant, C. E. (2017). Lipopolysaccharide-induced NF-κB nuclear translocation is primarily dependent on MyD88, but TNFα expression requires TRIF and MyD88. *Scientific Reports, 7*(1), 1428. doi:10.1038/s41598-017-01600-y

Sato, S., Sanjo, H., Takeda, K., Ninomiya-Tsuji, J., Yamamoto, M., Kawai, T., . . . Akira, S. (2005). Essential function for the kinase TAK1 in innate and adaptive immune responses. *Nature Immunology, 6*(11), 1087-1095. doi:10.1038/ni1255

Scaffidi, P., Misteli, T., & Bianchi, M. E. (2002). Release of chromatin protein HMGB1 by necrotic cells triggers inflammation. *Nature, 418*(6894), 191-195. doi:10.1038/nature00858

Scheibner, K. A., Lutz, M. A., Boodoo, S., Fenton, M. J., Powell, J. D., & Horton, M. R. (2006). Hyaluronan Fragments Act as an Endogenous Danger Signal by Engaging TLR2. *The Journal of Immunology, 177*(2), 1272. doi:10.4049/jimmunol.177.2.1272

Schulz, C., Gomez Perdiguero, E., Chorro, L., Szabo-Rogers, H., Cagnard, N., Kierdorf, K., . . . Geissmann, F. (2012). A lineage of myeloid cells independent of Myb and hematopoietic stem cells. *Science, 336*(6077), 86-90. doi:10.1126/science.1219179

Sedeno-Cortes, A. E., & Pavlidis, P. (2014). Pitfalls in the application of gene-set analysis to genetics studies. *Trends Genet, 30*(12), 513-514. doi:10.1016/j.tig.2014.10.001

Seok, J., Warren, H. S., Cuenca, A. G., Mindrinos, M. N., Baker, H. V., Xu, W., . . . Host Response to Injury, L. S. C. R. P. (2013). Genomic responses in mouse models poorly mimic human inflammatory diseases. *Proceedings of the National Academy of Sciences of the United States of America, 110*(9), 3507-3512. doi:10.1073/pnas.1222878110

Shalem, O., Sanjana, N. E., Hartenian, E., Shi, X., Scott, D. A., Mikkelsen, T. S., . . . Zhang, F. (2014). Genome-Scale CRISPR-Cas9 Knockout Screening in Human Cells. *Science, 343*(6166), 84. doi:10.1126/science.1247005

Shembade, N., Ma, A., & Harhaj, E. W. (2010). Inhibition of NF-kappaB signaling by A20 through disruption of ubiquitin enzyme complexes. *Science, 327*(5969), 1135-1139. doi:10.1126/science.1182364

Shen, S., Park, J. W., Lu, Z.-x., Lin, L., Henry, M. D., Wu, Y. N., . . . Xing, Y. (2014). rMATS: Robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences, 111*(51), E5593. doi:10.1073/pnas.1419161111

Shimazu, R., Akashi, S., Ogata, H., Nagai, Y., Fukudome, K., Miyake, K., & Kimoto, M. (1999). MD-2, a molecule that confers lipopolysaccharide responsiveness on Toll-like receptor 4. *Journal of Experimental Medicine, 189*(11), 1777-1782.

Shin, K.-J., Wall, E. A., Zavzavadjian, J. R., Santat, L. A., Liu, J., Hwang, J.-I., . . . Simon, M. I. (2006). A single lentiviral vector platform for microRNA-based conditional RNA interference and coordinated transgene expression. *Proceedings of the National Academy of Sciences, 103*(37), 13759-13764.

Silva, J. M., Mizuno, H., Brady, A., Lucito, R., & Hannon, G. J. (2004). RNA interference microarrays: High-throughput loss-of-function genetics in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America, 101*(17), 6548. doi:10.1073/pnas.0400165101

Smith, M. A., Choudhary, G. S., Pellagatti, A., Choi, K., Bolanos, L. C., Bhagat, T. D., . . . Starczynowski, D. T. (2019). U2AF1 mutations induce oncogenic IRAK4 isoforms and activate innate immune pathways in myeloid malignancies. *Nat Cell Biol, 21*(5), 640-650. doi:10.1038/s41556-019-0314-5

Song, D. H., & Lee, J.-O. (2012). Sensing of microbial molecular patterns by Toll-like receptors. *Immunological Reviews, 250*(1), 216-229. doi:10.1111/j.1600-065X.2012.01167.x

Stein, M., Keshav, S., Harris, N., & Gordon, S. (1992). Interleukin 4 potently enhances murine macrophage mannose receptor activity: a marker of alternative immunologic macrophage activation. *J Exp Med, 176*(1), 287-292. doi:10.1084/jem.176.1.287

Stelzer, G., Rosen, N., Plaschkes, I., Zimmerman, S., Twik, M., Fishilevich, S., . . . Lancet, D. (2016). The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Current Protocols in Bioinformatics, 54*(1), 1.30.31-31.30.33. doi:10.1002/cpbi.5

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., . . . Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A, 102*(43), 15545-15550. doi:10.1073/pnas.0506580102

Sun, J., Katz, S., Dutta, B., Wang, Z., & Fraser, I. D. (2017). Genome-wide siRNA screen of genes regulating the LPS-induced TNF-alpha response in human macrophages. *Sci Data, 4*, 170007. doi:10.1038/sdata.2017.7

Sun, J., Li, N., Oh, K. S., Dutta, B., Vayttaden, S. J., Lin, B., . . . Fraser, I. D. (2016). Comprehensive RNAi-based screening of human and mouse TLR pathways identifies species-specific preferences in signaling protein use. *Sci Signal, 9*(409), ra3. doi:10.1126/scisignal.aab2191

Sun, L., Wu, J., Du, F., Chen, X., & Chen, Z. J. (2013). Cyclic GMP-AMP synthase is a cytosolic DNA sensor that activates the type I interferon pathway. *Science, 339*(6121), 786-791.

Sung, M. H., Li, N., Lao, Q., Gottschalk, R. A., Hager, G. L., & Fraser, I. D. (2014). Switching of the relative dominance between feedback mechanisms in lipopolysaccharide-induced NF-kappaB signaling. *Sci Signal, 7*(308), ra6. doi:10.1126/scisignal.2004764

Sweet, M. J., & Hume, D. A. (1996). Endotoxin signal transduction in macrophages. *Journal of Leukocyte Biology, 60*(1), 8-26. doi:10.1002/jlb.60.1.8

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., . . . Bork, P. (2010). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research, 39*(suppl_1), D561-D568.

Takaesu, G., Kishida, S., Hiyama, A., Yamaguchi, K., Shibuya, H., Irie, K., . . . Matsumoto, K. (2000). TAB2, a novel adaptor protein, mediates activation of TAK1 MAPKKK by linking TAK1 to TRAF6 in the IL-1 signal transduction pathway. *Mol Cell, 5*(4), 649-658. doi:10.1016/s1097-2765(00)80244-0

Takao, K., & Miyakawa, T. (2015). Genomic responses in mouse models greatly mimic human inflammatory diseases. *Proc Natl Acad Sci U S A, 112*(4), 1167-1172. doi:10.1073/pnas.1401965111

Takeda, K., Kaisho, T., & Akira, S. (2003). Toll-Like Receptors. *Annual Review of Immunology, 21*(1), 335-376. doi:10.1146/annurev.immunol.21.120601.141126

Tannahill, G. M., Curtis, A. M., Adamik, J., Palsson-McDermott, E. M., McGettrick, A. F., Goel, G., . . . O'Neill, L. A. J. (2013). Succinate is an inflammatory signal that induces IL-1β through HIF-1α. *Nature, 496*, 238. doi:10.1038/nature11986
https://www.nature.com/articles/nature11986#supplementary-information

Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F. J., Del Risco, H., . . . Conesa, A. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. *Genome Res*. doi:10.1101/gr.222976.117

Tefferi, A., & Vardiman, J. W. (2009). Myelodysplastic Syndromes. *New England Journal of Medicine, 361*(19), 1872-1885. doi:10.1056/NEJMra0902908

Trinchieri, G. (1994). Interleukin-12: a cytokine produced by antigen-presenting cells with immunoregulatory functions in the generation of T-helper cells type 1 and cytotoxic lymphocytes. *Blood, 84*(12), 4008-4027. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/7994020

Tseng, G. C., Ghosh, D., & Feingold, E. (2012). Comprehensive literature review and statistical considerations for microarray meta-analysis. *Nucleic Acids Res, 40*(9), 3785-3799. doi:10.1093/nar/gkr1265

Tu, Z., Argmann, C., Wong, K. K., Mitnaul, L. J., Edwards, S., Sach, I. C., . . . Schadt, E. E. (2009). Integrating siRNA and protein–protein interaction data to identify an expanded insulin signaling network. *Genome Research, 19*(6), 1057-1067.

Ulevitch, R. J. (1993). Recognition of bacterial endotoxins by receptor-dependent mechanisms. In *Advances in immunology* (Vol. 53, pp. 267-289): Elsevier.

van Furth, R., & Cohn, Z. A. (1968). The origin and kinetics of mononuclear phagocytes. *J Exp Med, 128*(3), 415-435. doi:10.1084/jem.128.3.415
Velegraki, M., Papakonstanti, E., Mavroudi, I., Psyllaki, M., Tsatsanis, C., Oulas, A., . . . Papadaki, H. A. (2013). Impaired clearance of apoptotic cells leads to HMGB1 release in the bone marrow of patients with myelodysplastic syndromes and induces TLR4-mediated cytokine production. *Haematologica, 98*(8), 1206. doi:10.3324/haematol.2012.064642
Verstak, B., Stack, J., Ve, T., Mangan, M., Hjerrild, K., Jeon, J., . . . Mansell, A. (2014). The TLR signaling adaptor TRAM interacts with TRAF6 to mediate activation of the inflammatory response by TLR4. *Journal of Leukocyte Biology, 96*(3), 427-436. doi:10.1189/jlb.2A0913-487R
Visconte, V., Makishima, H., Jankowska, A., Szpurka, H., Traina, F., Jerez, A., . . . Tiu, R. V. (2012). SF3B1, a splicing factor is frequently mutated in refractory anemia with ring sideroblasts. *Leukemia, 26*(3), 542-545. doi:10.1038/leu.2011.232
Vivar, J. C., Pemu, P., McPherson, R., & Ghosh, S. (2013). Redundancy Control in Pathway Databases (ReCiPa): An Application for Improving Gene-Set Enrichment Analysis in Omics Studies and "Big Data" Biology. *OMICS: A Journal of Integrative Biology, 17*(8), 414-422. doi:10.1089/omi.2012.0083
Wang, L., Tu, Z., & Sun, F. (2009). A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila. *BMC Genomics, 10*(1), 220. doi:10.1186/1471-2164-10-220
Wang, L., Xi, Y., Sung, S., & Qiao, H. (2018). RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes. *BMC Genomics, 19*(1), 546. doi:10.1186/s12864-018-4932-2
Wang, M., & Marín, A. (2006). Characterization and prediction of alternative splice sites. *Gene, 366*(2), 219-227. doi:https://doi.org/10.1016/j.gene.2005.07.015
Wang, T., Wei, J. J., Sabatini, D. M., & Lander, E. S. (2014). Genetic Screens in Human Cells Using the CRISPR-Cas9 System. *Science, 343*(6166), 80. doi:10.1126/science.1246981
Wang, Y., Chen, T., Han, C., He, D., Liu, H., An, H., . . . Cao, X. (2007). Lysosome-associated small Rab GTPase Rab7b negatively regulates TLR4 signaling in macrophages by promoting lysosomal degradation of TLR4. *Blood, 110*(3), 962. doi:10.1182/blood-2007-01-066027
Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet, 10*(1), 57-63. doi:10.1038/nrg2484
Watanabe, T., Watanabe, S., & Kawaoka, Y. (2010). Cellular Networks Involved in the Influenza Virus Life Cycle. *Cell Host & Microbe, 7*(6), 427-439. doi:https://doi.org/10.1016/j.chom.2010.05.008
Wertz, I. E., O'Rourke, K. M., Zhou, H., Eby, M., Aravind, L., Seshagiri, S., . . . Dixit, V. M. (2004). De-ubiquitination and ubiquitin ligase domains of A20 downregulate NF-κB signalling. *Nature, 430*(7000), 694-699. doi:10.1038/nature02794
Whiteside, S. T., Epinat, J. C., Rice, N. R., & Israël, A. (1997). I kappa B epsilon, a novel member of the IκB family, controls RelA and cRel NF-κB activity. *The EMBO journal, 16*(6), 1413-1426.
Wickham, H. (2015). *R Packages*: O'Reilly Media.
Winston Chang, J. C., JJ Allaire, Yihui Xie and Jonathan McPherson. (2019). shiny: Web Application Framework for R. R package version 1.3.2. Retrieved from https://CRAN.R-project.org/package=shiny

Wright, S. D., Ramos, R. A., Tobias, P. S., Ulevitch, R. J., & Mathison, J. C. (1990). CD14, a receptor for complexes of lipopolysaccharide (LPS) and LPS binding protein. *Science, 249*(4975), 1431-1433.

Xiao, Y. Q., Malcolm, K., Worthen, G. S., Gardai, S., Schiemann, W. P., Fadok, V. A., . . . Henson, P. M. (2002). Cross-talk between ERK and p38 MAPK mediates selective suppression of pro-inflammatory cytokines by transforming growth factor-beta. *J Biol Chem, 277*(17), 14884-14893. doi:10.1074/jbc.M111718200

Xu, Y., Tao, X., Shen, B., Horng, T., Medzhitov, R., Manley, J. L., & Tong, L. (2000). Structural basis for signal transduction by the Toll/interleukin-1 receptor domains. *Nature, 408*(6808), 111.

Xue, J., Schmidt, S. V., Sander, J., Draffehn, A., Krebs, W., Quester, I., . . . Schultze, J. L. (2014). Transcriptome-based network analysis reveals a spectrum model of human macrophage activation. *Immunity, 40*(2), 274-288. doi:10.1016/j.immuni.2014.01.006

Yabas, M., Elliott, H., & Hoyne, G. F. (2015). The Role of Alternative Splicing in the Control of Immune Homeostasis and Cellular Differentiation. *Int J Mol Sci, 17*(1). doi:10.3390/ijms17010003

Yamamoto, M., Sato, S., Mori, K., Hoshino, K., Takeuchi, O., Takeda, K., & Akira, S. (2002). Cutting Edge: A Novel Toll/IL-1 Receptor Domain-Containing Adapter That Preferentially Activates the IFN-β Promoter in the Toll-Like Receptor Signaling. *The Journal of Immunology, 169*(12), 6668. doi:10.4049/jimmunol.169.12.6668

Yeo, G., & Burge, C. B. (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. *Journal of Computational Biology, 11*(2-3), 377-394. doi:10.1089/1066527041410418

Yona, S., Kim, K. W., Wolf, Y., Mildner, A., Varol, D., Breker, M., . . . Jung, S. (2013). Fate mapping reveals origins and dynamics of monocytes and tissue macrophages under homeostasis. *Immunity, 38*(1), 79-91. doi:10.1016/j.immuni.2012.12.001

Yoneyama, M., Kikuchi, M., Natsukawa, T., Shinobu, N., Imaizumi, T., Miyagishi, M., . . . Fujita, T. (2004). The RNA helicase RIG-I has an essential function in double-stranded RNA-induced innate antiviral responses. *Nat Immunol, 5*(7), 730-737. doi:10.1038/ni1087

Yoon, H. K., Byun, H. S., Lee, H., Jeon, J., Lee, Y., Li, Y., . . . Hur, G. M. (2013). Intron-derived aberrant splicing of A20 transcript in rheumatoid arthritis. *Rheumatology (Oxford), 52*(3), 427-437. doi:10.1093/rheumatology/kes292

Yu, D., Kim, M., Xiao, G., & Hwang, T. H. (2013). Review of biological network data and its applications. *Genomics & informatics, 11*(4), 200-210. doi:10.5808/GI.2013.11.4.200

Zhang, G., & Ghosh, S. (2002). Negative Regulation of Toll-like Receptor-mediated Signaling by Tollip. *Journal of Biological Chemistry, 277*(9), 7059-7065. Retrieved from http://www.jbc.org/content/277/9/7059

Zhang, H., Tay, P. N., Cao, W., Li, W., & Lu, J. (2002). Integrin-nucleated Toll-like receptor (TLR) dimerization reveals subcellular targeting of TLRs and distinct mechanisms of TLR4 activation and signaling. *FEBS letters, 532*(1-2), 171-176.

Zhang, J. H., Chung, T. D., & Oldenburg, K. R. (1999). A Simple Statistical Parameter for Use in Evaluation and Validation of High Throughput Screening Assays. *J Biomol Screen, 4*(2), 67-73. doi:10.1177/108705719900400206

Zhang, W., Chien, J., Yong, J., & Kuang, R. (2017). Network-based machine learning and graph theory algorithms for precision oncology. *npj Precision Oncology, 1*(1), 25. doi:10.1038/s41698-017-0029-7

Zhou, H., Xu, M., Huang, Q., Gates, A. T., Zhang, X. D., Castle, J. C., . . . Espeseth, A. S. (2008). Genome-scale RNAi screen for host factors required for HIV replication. *Cell Host Microbe, 4*(5), 495-504. doi:10.1016/j.chom.2008.10.004

Zhu, J., Davoli, T., Perriera, Jill M., Chin, Christopher R., Gaiha, Gaurav D., John, Sinu P., . . . Brass, Abraham L. (2014). Comprehensive Identification of Host Modulators of HIV-1 Replication using Multiple Orthologous RNAi Reagents. *Cell Rep, 9*(2), 752-766. doi:https://doi.org/10.1016/j.celrep.2014.09.031

Zhu, J., Mayeda, A., & Krainer, A. R. (2001). Exon Identity Established through Differential Antagonism between Exonic Splicing Silencer-Bound hnRNP A1 and Enhancer-Bound SR Proteins. *Molecular cell, 8*(6), 1351-1361. doi:https://doi.org/10.1016/S1097-2765(01)00409-9

Zhu, X., Santat, L. A., Chang, M. S., Liu, J., Zavzavadjian, J. R., Wall, E. A., . . . Fraser, I. D. C. (2007). A versatile approach to multiple gene RNA interference using microRNA-based short hairpin RNAs. *BMC molecular biology, 8*(1), 98.